arXiv:2412.16352v3 [econ.EM] 2 Oct 2025

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect *

Neil Christy† and Amanda Ellen Kowalski‡

October 1, 2025

Abstract

Leveraging structure from the randomization process, a design-based model of an experiment with a binary intervention and outcome can reveal evidence beyond the average effect without additional data. Our proposed statistical decision rule yields a design-based maximum likelihood estimate (MLE) of the joint distribution of potential outcomes in intervention and control, specified by the numbers of always takers, compliers, defiers, and never takers in the sample. With a visualization, we explain why the likelihood varies with the number of defiers within the Fréchet bounds determined by the estimated marginal distributions. We illustrate how the MLE varies with all possible data in samples of 50 and 200: when the estimated average effect is positive, the MLE includes defiers if takeup is below half in control and above half in intervention, unless takeup is zero in control or full in intervention. Under optimality conditions, for increasing sample sizes in which exhaustive grid search is possible, our rule's performance increases relative to a rule that places equal probability on all numbers of defiers within the estimated Fréchet bounds. We offer insights into effect heterogeneity in two published experiments with positive, statistically

*This paper is the culmination of a series of drafts (Kowalski, 2019a,b; Christy and Kowalski, 2024a,b). We extend special thanks to Jann Spiess for extensive regular feedback and to Charles Manski, Aleksey Tetenov, Toru Kitagawa, and Donald Rubin for encouraging us to use statistical decision theory and teaching us about it. We thank Guido Imbens for foundational feedback. Tory Do, Simon Essig Aberg, Jack Cavanaugh, Bailey Flanigan, Pauline Mourot, Srajal Nayak, Darion Phan, Sukanya Sravasti, Matthew Tauzer, and Shuheng Zhang provided excellent research assistance. For helpful comments, we thank Elizabeth Ananat, Don Andrews, Isaiah Andrews, Josh Angrist, Susan Athey, Victoria Baranov, Steve Berry, Stephane Bonhomme, Michael Boskin, Zach Brown, Kate Bundorf, Matias Cattaneo, Xiaohong Chen, Victor Chernozhukov, Janet Currie, Peng Ding, Pascaline Dupas, Brad Efron, Natalia Emanuel, Ivan Fernandez-Val, Chris Ferrie, Ashvin Gandhi, Michael Gechter, Andrew Gelman, Matthew Gentzkow, Sander Greenland, Florian Gunsilius, Andreas Hagemann, Sukjin Han, Jerry Hausman, Han Hong, Lawrence Katz, Daniel Kessler, Michal Kolesár, Jonathan Kolstad, Ang Li, John List, Bentley MacLeod, Aprajit Mahajan, José Luis Montiel Olea, Sendhil Mullainathan, Derek Neal, Andriy Norets, Matthew Notowidigdo, Elena Pastorino, John Pepper, Demian Pouzo, Tanya Rosenblat, Azeem Shaikh, Elie Tamer, Davide Viviano, Edward Vytlacil, Stefan Wager, Chris Walker, Christopher Walters, Thomas Wiemann, David Wilson, seminar participants at Columbia, Harvard, MIT, the Norwegian School of Economics, Notre Dame, NYU, Princeton, Stanford, UCLA, UVA, the University of Chicago, the University of Michigan, the University of Pennsylvania, the University of Zurich, Yale, and conference participants at the Advances with Field Experiments Conference at the University of Chicago, the AEA meetings, the Bravo Center/SNSF Workshop on Using Data to Make Decisions, the Essen Health Conference, and the Y-RISE Evidence Aggregation and External Validity Conference. We thank Charles Antonelli, Bennett Fauber, Corey Powell, and Advanced Research Computing at the University of Michigan, as well as Misha Guy, Andrew Sherman, and the Yale High Performance Computing Center. This research was supported in part by the National Institute on Aging of the National Institutes of Health under Award Number R01AG089084. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

†University of Michigan

‡University of Michigan and the National Bureau of Economic Research

significant average effects on takeup of desired health behaviors and plausible defiers. Our 95% smallest credible sets for defiers include zero and the estimated upper Fréchet bound, demonstrating that evidence is weak. Yet, our rule yields no defiers in one experiment. In the other, our rule yields the estimated upper Fréchet bound on defiers—a count representing over 18% of the sample.

Keywords: econometrics, statistical decision theory, experiments, health
JEL Codes: C1,C44,C9, I1

1 Introduction

We begin with a stylized example that demonstrates how a design-based model of an experiment reveals evidence beyond the average effect. In Section 2, we formalize the model and present a likelihood function for the joint distribution of potential outcomes in the sample, which we use to construct a maximum likelihood decision rule. In Section 3, we apply our maximum likelihood decision rule to two real-world experiments. Section 4 discusses our contributions to the literature, which we reserve until the end for clarity. Section 5 presents implications for econometric and applied research.

1.1 Visualization of a Stylized Example: A Design-Based Model Can Reveal Evidence Beyond the Average Effect

Suppose you are considering an intervention to increase takeup of a desired health behavior. For example, the desired health behavior could be to quit smoking or to become an organ donor, and the intervention could be a payment or a behavioral nudge. You want to estimate the average effect of the intervention. You would also like to decide if the intervention backfires, and if so, for how many people. Your colleague tells you that an experiment can be informative about the population average effect of the intervention, but it cannot provide evidence that the intervention backfires if the estimated average effect is positive. Instead of considering the population, you focus your attention on effects within the fixed set of subjects in an experiment.

You design an experiment with six people such that you randomly assign exactly three people to intervention and three people to control. You run the experiment, and you estimate the marginal distributions of outcomes in intervention and control: two people have taken up the behavior in intervention, and one person has taken up the behavior in control. These results seem promising because takeup is higher in intervention than control, yielding an average effect of 1/3 on a control takeup rate of 1/3.

Did the intervention backfire for anyone? *Why* did you observe the data you observed? What would you have seen under the other possible randomized assignments? In more technical terms, what is the joint distribution of "potential outcomes" (Neyman, 1923; Rubin, 1974, 1977) in the sample? Counterfactual questions like these have attracted recent interest in the study of causal inference (Gelman and Imbens, 2013; Pearl and Mackenzie, 2018; Imbens, 2020; Dawid and Musio, 2022).

We assume as in design-based inference that the potential outcomes of each person are fixed, and therefore the joint distribution of potential outcomes in the sample is fixed, before randomization. Following Angrist, Imbens, and Rubin (1996), we classify people into types based on their potential outcomes in intervention and control: "always takers" take up in intervention and control (effect=0), "compliers" take up in intervention but not control (effect=1), "defiers" do not take up in intervention but take up in control (effect=-1), and "never takers" do not take up in intervention or control (effect=0). We specify the joint distribution of potential outcomes

by enumerating the numbers of always takers, compliers, defiers, and never takers in the sample. The joint distribution of potential outcomes provides more information than the average effect; it provides the full distribution of effects and disambiguates between zero effects from always and never takers. What is your best guess of the numbers of always takers, compliers, defiers, and never takers in the experiment you ran with six people?

We decide with our design-based maximum likelihood decision rule that the sample includes four compilers and two defiers, and we develop a novel visualization of potential outcomes to illustrate the logic behind our answer in Figure 1. We represent each of the six people in the sample with a circle. The left half of each circle represents the potential outcome in intervention and the right half represents the potential outcome in control. Orange represents "takeup," white represents "no takeup," and grey represents "unobserved."

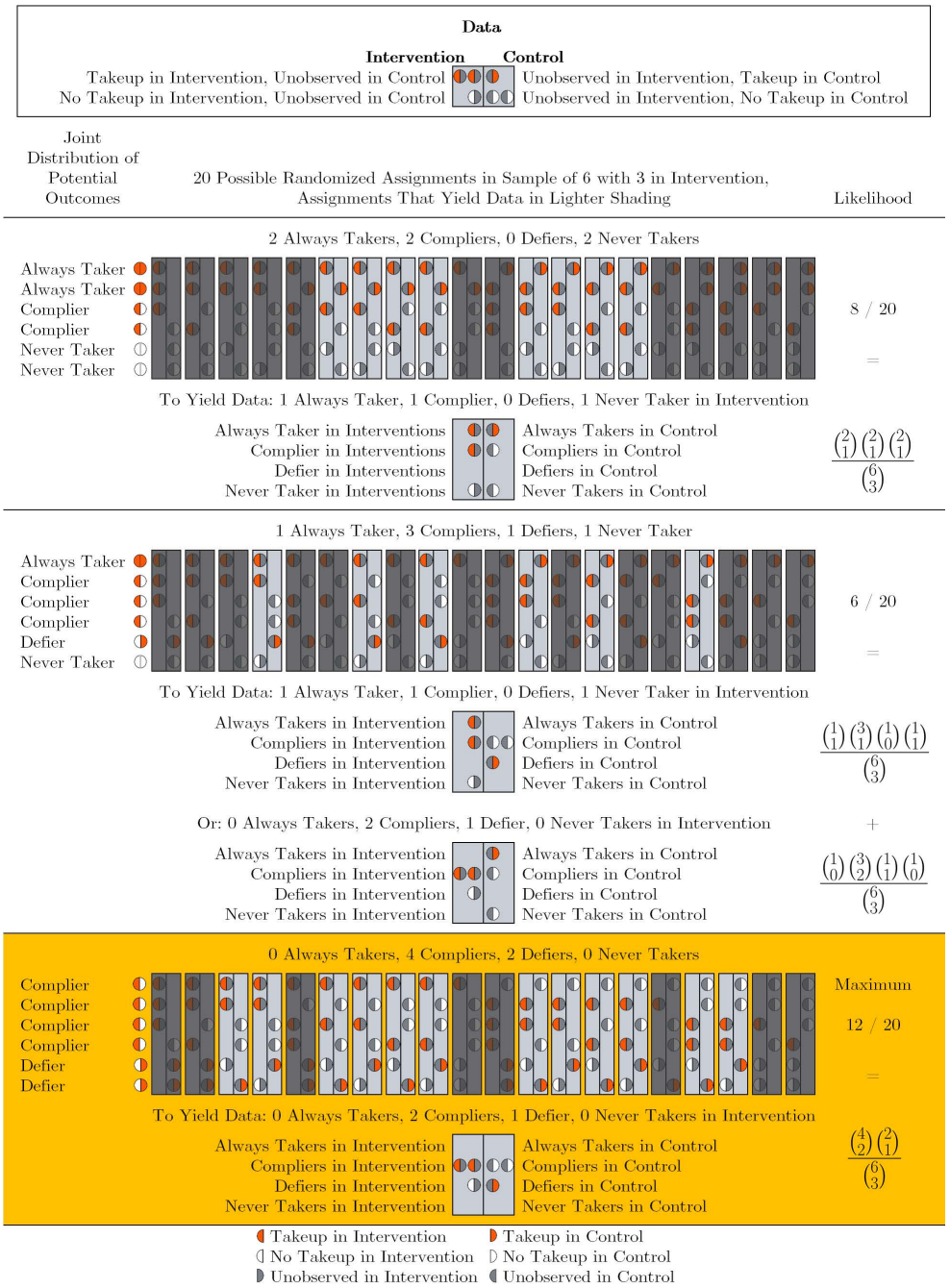
Rectangles with lines down the middle represent experiments with intervention on the left and control on the right. Inside each experiment, assignment to intervention reveals the potential outcome in intervention—the left of the rectangle reveals the left of the circle, which can be orange for takeup or white for no takeup—but the potential outcome in control is unobserved, so it is grey. Similarly, assignment to control reveals the potential outcome in control–the right of the rectangle reveals the right of the circle, which can be orange for takeup or white for no takeup—but the potential outcome in intervention is unobserved, so it is grey.

The data from your experiment, depicted in the box at the top of Figure 1 shows takeup among two of three people in intervention and one of three people in control. Therefore, the estimated marginal distribution of the potential outcome in intervention is 2/3 takeup, and the estimated marginal distribution of the potential outcome in control is 1/3 takeup. These two marginal distributions define the "estimated Fréchet set." We refer to all joint distributions of potential outcomes with the same marginal distributions as members of the same "Fréchet set."

The three rows of Figure 1 give the three joint distributions of potential outcomes with the same marginal distributions as the estimated Fréchet set: half circles in which the potential outcomes are unobserved in the data are filled in so that they have the same marginal distributions as the data. Takeup occurs among 2/3 of people in intervention (four of the six circles have orange on the left) and among 1/3 of people in control (two of the six circles have orange on the right). As shown in the three rows, there are three ways to combine the colored half circles into colored full circles, each yielding a different number of defiers (circles with white on the left and orange on the right). This depiction demonstrates how the estimated Fréchet set determines the estimated Fréchet bounds on defiers, which extend here from a lower bound of 0 to an upper bound of 2. How do you decide among these three answers?

The joint distribution of potential outcomes in the first row is consistent with the Imbens and Angrist (1994) assumption of "monotonicity" in the sample, whereby there can be compliers or defiers, but not both. Since the estimated effect of 1/3

Figure 1: In an Experiment with Six People, the Likelihood Varies Among Joint Distributions of Potential Outcomes with the Same Marginal Distributions as the Data (2/3 Takeup in Intervention, 1/3 Takeup in Control) and is Maximized with Four Compilers and Two Defiers



is positive, you assume away defiers. You then deduce that the person who takes up in control must be an always taker and the person who does not take up in intervention must be a never taker. Assuming that the randomization guarantees that the count of each type is the same in each arm, the sample includes two always takers and two never takers. The remaining two people must be compliers, so the joint distribution of potential outcomes consistent with the monotonicity assumption includes two always takers, two compliers, and two never takers. However, you are uncomfortable assuming monotonicity because you are concerned that there could be defiers, and you would prefer to decide based on evidence.

The joint distributions of potential outcomes in the next two rows include defiers. The one in the second row includes one always taker, three compliers, one defier, and one never taker. The one in the third row includes four compliers and two defiers. There is a rationale in the literature for deciding on the second row—it has one defier, the midpoint (Li and Pearl, 2019) of the Fréchet bounds on defiers. It can therefore seem surprising that we decide on the distribution in the third row.

The top half of each row in Figure 1 provides a rationale for our decision using a "design-based model" that combines the design of the randomization with a model of potential outcomes to specify a data generating process and use it to generate all possible data. As depicted in the top of each row, there are 20 possible randomized assignments of six people that result in three in intervention (6 choose 3 = 20). Each of the six people in the experiment has a separate row. The person's type and randomized assignment to intervention or control determines the data the experiment reveals, which can be counterfactual to what you observed. The assignments that yield the data you observed—2/3 takeup in intervention and 1/3 takeup in control—are in lighter shading.

The number of randomized assignments that yield the data you observed gives the numerator of the design-based likelihood, reported in the third column. Paraphrasing the baby board book "Statistical Physics for Babies" (Ferrie, 2017), which provides style inspiration for our visualization, physicists refer to the number of ways that you could have seen what you have seen as "entropy." The design-based likelihood is equal to the entropy—the number of assignments that yield the data (shown in lighter shading)—divided by the number of possible assignments. The first row, with zero defiers, yields the data in eight of 20 possible assignments, so the likelihood is $8/20 = 40\%$. The second row, with one defier, yields the data in six of 20 possible assignments, so the likelihood is $6/20 = 30\%$. The last row, with two defiers, yields the data in 12 of 20 possible assignments, so the likelihood is $12/20 = 60\%$.

Why do some joint distributions of potential outcomes yield the data in more randomized assignments than others and therefore have higher likelihoods than others? The bottom half of each row sorts the six people by type and separately shows how the always takers, compliers, defiers, and never takers must be randomized into intervention and control to yield the data. In the first row, there are three types—always takers, compliers, and never takers—and there must be a balance between

intervention and control within each type to yield the data. In the second row, there are four types—always takers, compliers, defiers, and never takers—and there must be an imbalance between intervention and control within each type to yield the data. In the third row, there are only two types—compliers and defiers—and there must be a balance between intervention and control within each type to yield the data.

People can be the same or different, but the likelihoods are higher when the people who yield the data are as similar as possible in two senses: they belong to fewer types, and those of the same type are balanced between intervention and control.[1] The people who yield the data are most similar in the third row: they belong to the fewest number of types (only two), and those of the same type are balanced between intervention and control. Comparison of the first and third rows illustrates how having fewer types can increase the likelihood. In the first row, there two ways to randomize one of two always takers into intervention (2 choose $1 = 2$), two ways to randomize one of two compliers into intervention (2 choose $1 = 2$), and two ways to randomize one of two defiers into intervention (2 choose $1 = 2$), so there are eight assignments that yield the data ($2 \times 2 \times 2 = 8$). In the third row, there are six ways to randomize two of four compliers into intervention (4 choose $2 = 6$) and two ways to randomize one of two defiers into intervention (2 choose $1 = 2$), so there are 12 assignments that yield the data ($6 \times 2 = 12$). All assignments that yield the data in the first row also yield the data in the third row, which has fewer types. In the third row, having four compliers makes it possible for four additional assignments—assignments in which the first two people are in the same arm—to yield the data. Joint distributions of potential outcomes other than the ones in Figure 1 can yield the data, but by exhaustive grid search, you determine that the global maximizer of the likelihood is the one depicted in the third row, which only includes two types.

By our proposed maximum likelihood decision rule, you conclude that your best guess for the joint distribution of potential outcomes in the sample consists of four compilers and two defiers. Using this distribution, you recover the estimated average effect of 1/3—of the six people in the experiment, the four compliers have an effect of positive one, and the two defiers have an effect of negative one ($(4 \times 1 + 2 \times -1)/6 = 1/3$). And, while you know the experimental data provides no information about the population share for whom the experiment would backfire, you are able to estimate that, within your sample, the intervention backfired for two subjects. The design-based model of the experiment enables you to learn about not just the average effect of your intervention, but also about something more primitive: the joint distribution of potential outcomes, which yields the entire distribution of effects.

[1] Andrew Gelman and Keith O-Rourke discuss the importance of "similarity" in statistical evidence: "Awareness of commonness can lead to an increase in evidence regarding the target; disregarding commonness wastes evidence; and mistaken acceptance of commonness destroys otherwise available evidence. It is the tension between these last two processes that drives many of the theoretical and practical controversies within statistics" (Gelman and O'Rourke, 2017).

2 A Design-Based Model of an Experiment

2.1 A Design-Based Model Implies a Likelihood Function for the Joint Distribution of Potential Outcomes in the Sample

In this section, we formalize the design-based model of an experiment presented in our stylized example. A sample consists of n subjects, each of whom has a binary potential outcome in intervention $Y_I \in \{0, 1\}$ and a binary potential outcome in control $Y_C \in \{0, 1\}$, where 1 represents takeup and 0 represents no takeup. A subject's observed outcome depends only on their own potential outcomes and their inclusion in the intervention or control arm, ruling out network-type effects through a "no interference" (Cox, 1958) or "stable unit treatment value" (Rubin, 1980) assumption.

Subjects are randomly assigned to intervention ($Z = I$) or control ($Z = C$), and assignment to intervention Z is independent of a subject's potential outcomes (Y_I, Y_C). The observed outcome Y is:

$$Y = \mathbf{1}_{\{Z=I\}}(Y_I) + \mathbf{1}_{\{Z=C\}}(Y_C), \tag{1}$$

where $\mathbf{1}_{\{\bullet\}}$ is the indicator function.

Subjects in the sample belong to one of four types or "principal strata," defined by combinations of potential outcomes (Frangakis and Rubin, 2002). Let θ_{11} represent the number of always takers ($Y_I = 1, Y_C = 1$), θ_{10} the number of compilers ($Y_I = 1, Y_C = 0$), θ_{01} the number of defiers ($Y_I = 0, Y_C = 1$), and θ_{00} the number of never takers ($Y_I = 0, Y_C = 0$). The collection of these four counts $\boldsymbol{\theta} = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ constitutes the joint distribution of potential outcomes in the sample, which we write as counts rather than shares to ease the following likelihood notation. The experimental data $\boldsymbol{X} = (X_{I1}, X_{I0}, X_{C1}, X_{C0})$ consist of the counts of subjects who take up in intervention X_{I1}, who do not take up in intervention X_{I0}, who take up in control X_{C1}, and who do not take up in control X_{C0}.

We adopt a "design-based" approach, in which the joint distribution of potential outcomes $\boldsymbol{\theta}$ is fixed, but unknown, and all randomness in the experimental data $\boldsymbol{X}$ comes from the random assignment of subjects to either intervention or control. As shown in Appendix B, randomization implies a distribution for the experimental data $\boldsymbol{X}$ given the joint distribution of potential outcomes $\boldsymbol{\theta}$, yielding a likelihood expression. For Bernoulli randomization, also known as simple randomization or independent and identically distributed (IID) or randomization, with probability p of

assignment to intervention, the likelihood is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \\ & \times p^{x_{I1} + x_{I0}} (1-p)^{x_{C1} + x_{C0}},\end{aligned}$$

where the set $\mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})$ restricts i such that the binomial coefficients remain well defined.

For a completely randomized experiment with m subjects randomized into intervention, the likelihood is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{m + x_{C1} + i - \theta_{11} - \theta_{01} - x_{I1}} \Big/ \binom{n}{m}\end{aligned}$$

While these randomization procedures imply different distributions for the experimental data, once we condition on the experimental data, the likelihoods are both proportional to:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto \sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{m + x_{C1} + i - \theta_{11} - \theta_{01} - x_{I1}}\end{aligned} \tag{2}$$

This last expression appears in Copas (1973).

This likelihood function coincides with the likelihoods we derive in the introductory example of Section 1. For the first distribution with zero defiers and the third distribution with two defiers in Figure 1, there is a single term in the likelihood summation. For the second distribution with one defier, there are two.

2.2 The Design-Based Likelihood Can Vary with the Number of Defiers within the Fréchet Bounds Determined by the Estimated Marginal Distributions of Potential Outcomes

In the previous section, we present a likelihood for the joint distribution of potential outcomes in the sample. Each joint distribution of potential outcomes implies marginal distributions of potential outcomes in the intervention and control arms. The marginal distribution of the potential outcome in intervention is:

$$\theta_{1\bullet} \equiv \theta_{11} + \theta_{10}.$$

The marginal distribution of the potential outcome in control is:

$$\theta_{\bullet 1} \equiv \theta_{11} + \theta_{01}.$$

Given a pair of marginal distributions of potential outcomes in intervention and control, there are numerous joint distributions consistent with them. Boole (1854), Hoeffding (1940), and Fréchet (1957) derive bounds on the possible copulas connecting marginal distributions of random variables into joint distributions. By applying these bounds to a pair of marginal distributions of potential outcomes, we can determine the set of all joint distributions of potential outcomes consistent with the pair of marginal distributions. We refer to the set of joint distributions of potential outcomes consistent with a given pair of marginal distributions as the "Fréchet set" of those marginal distributions, $\mathcal{F}(\theta_{1\bullet}, \theta_{\bullet 1}; n)$. We index the distributions in a Fréchet set by their number of defiers, θ_{01}, which can take values in the range

$$\max\{0, -(\theta_{1\bullet} - \theta_{\bullet 1})\} \leq \theta_{01} \leq \min\{\theta_{\bullet 1}, n - \theta_{1\bullet}\} \tag{3}$$

These inequalities provide the so-called "Fréchet bounds" on the number of defiers. Within these bounds, each value for the number of defiers implies a unique number of always takers, never takers, and compilers consistent with the given marginals:

$$\begin{aligned}\mathcal{F}(\theta_{1\bullet}, \theta_{\bullet 1}; n) = \Big\{ \boldsymbol{\theta} \in \mathbb{Z}_+^4 : & \theta_{01} \in [\max\{0, -(\theta_{1\bullet} - \theta_{\bullet 1})\}, \min\{\theta_{\bullet 1}, n - \theta_{1\bullet}\}], \\ & \theta_{10} = \theta_{1\bullet} - \theta_{\bullet 1} + \theta_{01}, \\ & \theta_{11} = \theta_{\bullet 1} - \theta_{01}, \\ & \theta_{00} = n - \theta_{1\bullet} - \theta_{01} \Big\}\end{aligned}$$

The bounds on the number of defiers θ_{01} ensure that these counts are non-negative

and sum to n.

We can form estimates of the marginal distributions of potential outcomes in intervention and control directly using the observed takeup rates in intervention and control:

$$\hat{\theta}_{1\bullet}(\boldsymbol{x}) = \Big(\frac{n}{\mathbb{P}(\text{Assignment to Intervention})}\Big)\Big(\frac{x_{I1}}{x_{I1}+x_{I0}}\Big)$$
$$\hat{\theta}_{\bullet 1}(\boldsymbol{x}) = \Big(\frac{n}{\mathbb{P}(\text{Assignment to Control})}\Big)\Big(\frac{x_{C1}}{x_{C1}+x_{C0}}\Big). \tag{4}$$

When we report these estimates, we round them if necessary for comparison to our count-based distributions. In our design-based model of an experiment, these are unbiased estimators of the sample marginal distributions of potential outcomes (before rounding). In a standard sampling-based model of an experiment, these estimators are both unbiased and consistent estimates of the population marginal distributions of potential outcomes.

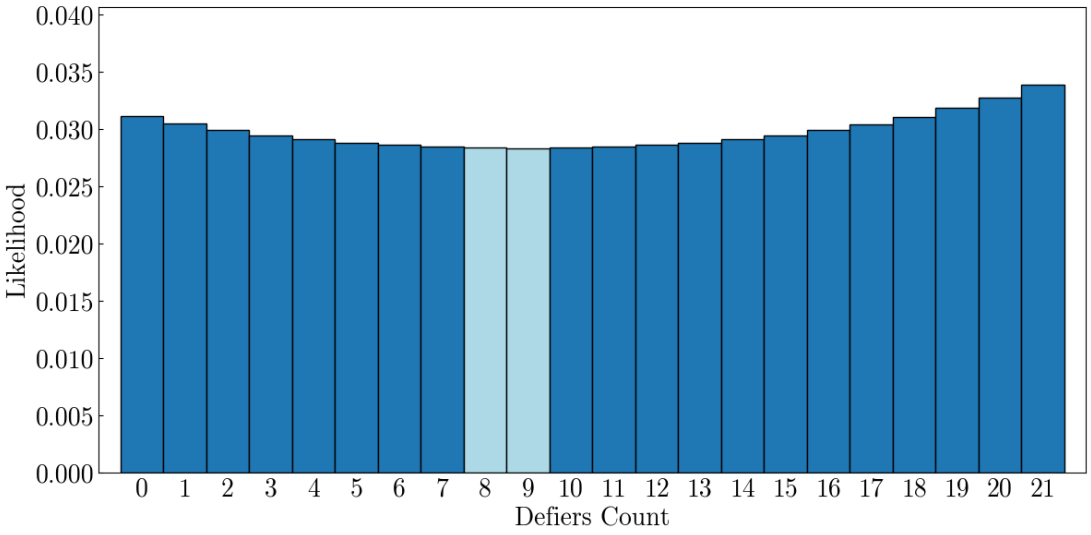
It is well-known that, in the sampling-based model of an experiment, the takeup counts in intervention and control contain no information about the joint distribution of potential outcomes in the population beyond their marginal distributions. In Appendix C, we recreate this result by showing that the likelihood of the population-level joint distribution of potential outcomes is constant, conditional on the population-level marginal distributions. In other words, the likelihood function is flat across every population-level Fréchet set.

In our design-based model of an experiment, however, the data does contain information about the sample joint distribution of potential outcomes beyond the sample marginal distributions. That is, the likelihood function varies with the number of defiers within sample-level Fréchet sets. Figure 2 plots the likelihood values of each joint distribution of potential outcomes in the Fréchet set formed by the estimated marginal distributions from the second empirical example that we will discuss in Section 3. We index the joint distributions in this Fréchet set by their number of defiers along the horizontal axis. The height of the bar is the likelihood of the joint distribution in the Fréchet set with the specified number of defiers. Within this Fréchet set, the likelihood is highest at the distribution with 21 defiers, the upper Fréchet bound on defiers. As the number of defiers increases, the likelihood initially decreases, and then increases. In all empirical examples we have considered, the likelihood maintains a U-shape within the estimated Fréchet set and all other Fréchet sets, and is maximal at either the minimum or maximum number of defiers. If we normalize likelihoods in this Fréchet set to sum to one, the smallest collection of likelihoods to contain 95% of the mass excludes 8 and 9 defiers, indicated by their lighter shading.

2.3 Proposed Maximum Likelihood Decision Rule

We consider a statistical decision problem in the style of Wald (1949), whereby the decision maker wishes to correctly guess the joint distribution of potential outcomes

Figure 2: Likelihood Varies with the Number of Defiers within the Estimated Fréchet Set in Johnson and Goldstein (2003), Smallest Collection of Likelihoods to Contain 95% of the Mass Excludes 8 and 9 Defiers (Shown in Lighter Shading)



in the sample, having viewed the outcome of the experiment. In Appendix D.1, we present the formal decision problem. Statistical decision theory provides two benefits in our setting: first, decision theory is natural to apply in our design-based setting, unlike alternative criteria like consistency that depend on large-sample or asymptotic assumptions; second, statistical decision theory provides straightforward methods to quantify the gains from exploiting the variation in our likelihood over other decision rules that ignore such variation.

We propose choosing the joint distribution of potential outcomes that maximizes the design-based likelihood. Let $\widehat{\boldsymbol{\Theta}}(\boldsymbol{x})$ be the set of joint distributions $\boldsymbol{\theta}$ that maximize the likelihood given the observed data $\boldsymbol{x}$:

$$\widehat{\boldsymbol{\Theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$$

There are finitely many vectors of four integers that sum to the number of participants in the experiment n, so $\widehat{\boldsymbol{\Theta}}(\boldsymbol{x})$ is nonempty. We define the maximum likelihood decision rule as an estimator $\hat{\boldsymbol{\theta}}_{MLE}$ that places equal weight on each maximizer of the likelihood:

$$\mathbb{P}(\hat{\boldsymbol{\theta}}_{MLE}(\boldsymbol{x}) = \boldsymbol{\theta} \mid \boldsymbol{x}) = \begin{cases} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}(\boldsymbol{x})\}}, & \text{if} \quad \boldsymbol{\theta} \in \widehat{\boldsymbol{\Theta}}(\boldsymbol{x}), \\ 0, & \text{otherwise} \end{cases}$$

where $\#\{\bullet\}$ is the counting measure. When the likelihood is unimodal, the decision

rule chooses the maximizer with probability one. We solve this optimization problem through an exhaustive grid search over the finitely many joint distributions of potential outcomes consisting of n subjects.

2.4 Justifications for our Proposed Maximum Likelihood Decision Rule

Our primary motivation for proposing the maximum likelihood decision rule is that maximum likelihood rules are considered "standard" in econometric settings, and maximum likelihood estimators are familiar to practitioners. In the following subsections, we provide three additional justifications: first, that the maximum likelihood decision rule varies systematically with the data; second, that the decision rule is Bayes optimal under the appropriate conditions; and third, that the decision rule maximizes entropy, and is therefore the "least informative" distribution to justify the data.

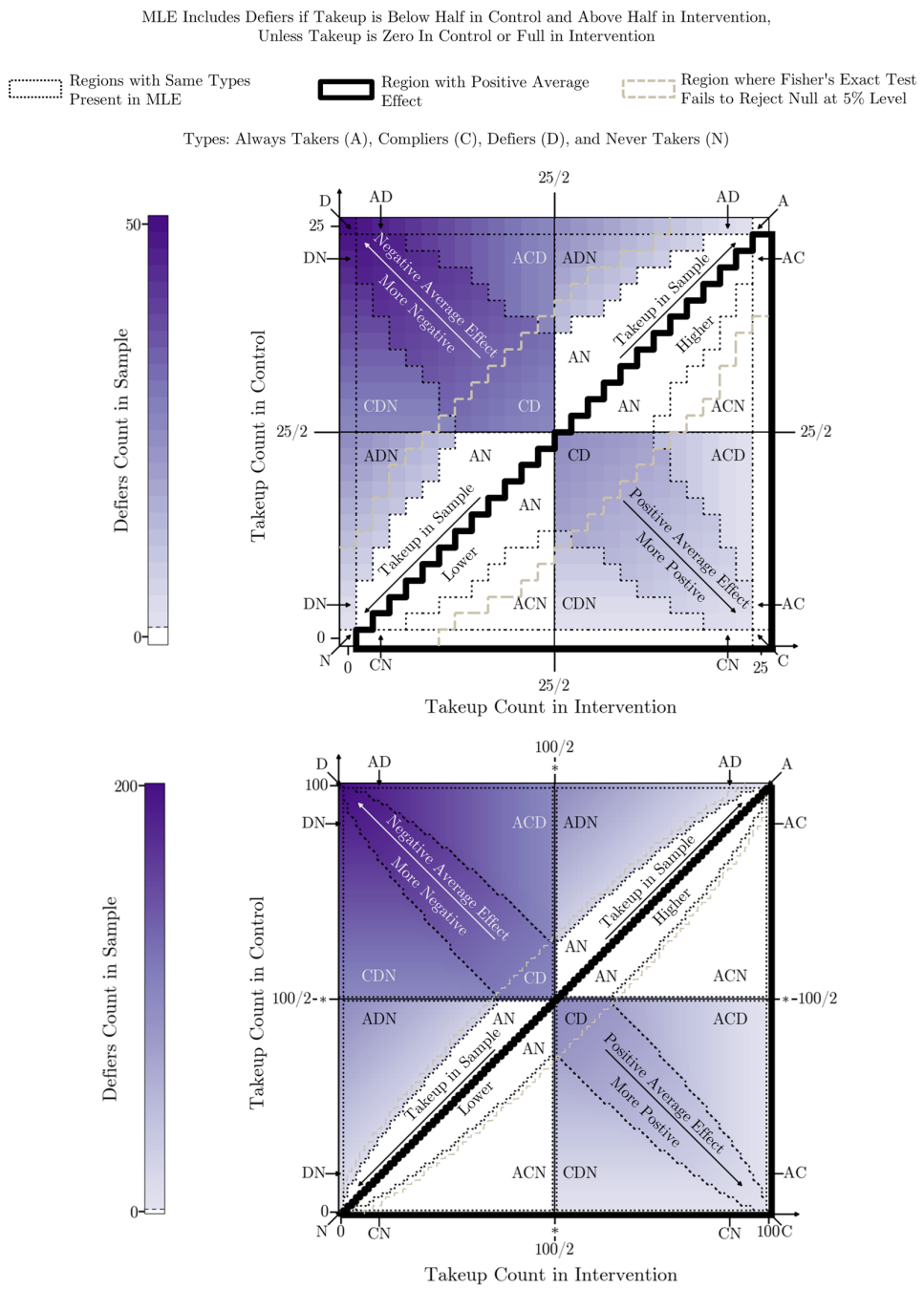
2.4.1 The Maximum Likelihood Decision Rule Varies Systematically with the Data

Direct inspection of the likelihood in (2) does not reveal obvious patterns in the maximum likelihood estimate. However, by computing the maximum likelihood estimate for every possible realization of the data in fixed sample sizes, we can observe clear patterns in the estimates. Figure 3 illustrates how the maximum likelihood estimates from our proposed decision rule vary systematically with all possible realizations of the data in completely randomized experiments with 50 and 200 people such that half are in intervention. In the subfigure for each sample size, the horizontal axis gives the takeup count in intervention, and the vertical axis gives the takeup count in control. Each cell represents a possible realization of the data, which specifies an estimated Fréchet set in terms of takeup count in intervention and control. The visualization demonstrates that an estimated Fréchet set can also be specified with the estimated average effect, which increases diagonally downward, and the total takeup count in the sample, which increases diagonally upward. An estimated Fréchet set can also be specified with an estimated average effect and a takeup rate in control, as is common in empirical practice.

The figure provides a visualization of our statistical decision rule by reporting statistics from our maximum likelihood estimates of the numbers of always takers, compliers, defiers, and never takers in the sample. Maximum likelihood estimates in cells within dotted outlines include the same types: always takers (A), compliers (C), defiers (D), and never takers (N). Darker purple shading within cells indicates higher counts of defiers in the maximum likelihood estimates.

The figure shows clear patterns in how the MLEs vary with the data, which are visibly similar when we increase the sample size by a factor of four. MLEs in the corners only include one type: always takers if takeup is full in intervention and control, compliers if takeup is full in intervention and zero in control, defiers if takeup is zero in intervention and full in control, and never takers if takeup is zero in intervention and control. MLEs along the edges only include the two types in the

Figure 3: Visualization of Our Proposed Statistical Decision Rule: How the MLE Varies with All Possible Data from Samples of 50 and 200 with Half in Intervention



* Two-way ties in the MLE occur in cells bisected by indicated lines, between MLEs in the cells on either side of the line, only when takeup count in intervention and control is not zero or full.

corners that they connect.

In the lens-shaped region around the upward-sloping diagonal, MLEs only include always and never takers: the two types in the null hypothesis of the Fisher (1935) exact test. In samples of 50 and 200, the region in which Fisher's exact test fails to reject the null hypothesis at the 5% level is depicted within the grey dashed line. It is larger than the region in which the MLE includes only always and never takers in both sample sizes.

The figure illustrates how our design-based maximum likelihood decision rule reveals evidence that is obscured by Fisher's exact test. Fisher's exact test only considers the null hypothesis that the joint distribution of potential outcomes includes only always and never takers. If it rejects, it does not suggest an alternative. If it fails to reject, a joint distribution of potential outcomes that include types other than always and never takers can maximize the likelihood. With our use of statistical decision theory, we can harness the richer evidence conveyed in the figure beyond the grey dashed lines.

In the lens-shaped region around the downward sloping diagonal, MLEs only include compliers and defiers. This region is related to the region in which an exact test of the null hypothesis that everyone is affected does not reject. There are some realizations of the data near the center of the region in which Fisher's exact test fails to reject the null hypothesis that no one is affected, but our maximum likelihood estimates indicate that everyone is affected. These estimates are relevant to empirical practice for researchers who wonder if they obtained a near-zero estimated average effect not because no one was affected but because everyone was affected.

Sometimes the MLE can be a tie between more than one distribution. MLEs are unique in the sample of 50, but two-way ties occur in the sample of 200. In the cells bisected by the lines indicated with asterisks–the midlines in which takeup is exactly half in intervention or control–ties occur between the MLEs in the cells on either side of the line. Only at the edges of the figure do the ties result in a unique MLE because the MLE is the same on either side of the line.

The MLEs include three types in all other regions, two regions for each combination. Even though the maximizer of the likelihood only includes two types in our stylized example from Figure 1, maximization of the likelihood sometimes requires trading off between the higher likelihood of fewer types and the higher likelihood of more balance between intervention and control within each type. If restricting the sample to two types requires too much imbalance between intervention and control, the maximizer of the likelihood includes three types.[2] There are no regions in which a distribution that maximizes the likelihood contains all four types. Therefore, many joint distributions of potential outcomes never maximize the likelihood for any

[2]Viewing Pascal's triangle as a representation of N choose K can help us to visualize the tradeoff between fewer types (moving to the next row of the triangle, increasing N within each type) and more balance within each type (moving to the middle of the triangle within a row, moving K closer to $N/2$ within each type). We thank Elizabeth Ananat for this point.

realization of the data.

The region outlined in bold, in which the estimated average effect is positive, is of particular interest because MLEs in the shaded portion of the region include both defiers and compliers, contrary to monotonicity assumptions, as in the stylized example in Figure 1. In all even-sized samples up to 200, the MLE includes defiers when takeup is below half in control and above half in intervention, unless takeup is zero in control or full in intervention. The experiment depicted in Figure 1 falls into the region with defiers because the takeup rate is 1/3 in control and 2/3 in intervention. As shown, the MLE always includes defiers when the estimated average effect is greater than half, unless takeup is zero in control or full in intervention, and it can include defiers even when Fisher's exact test rejects at the 5% level.

The shading demonstrates that the MLE reveals evidence beyond the average effect: holding the estimated average effect constant along upward sloping diagonal lines, MLEs can include zero or many defiers. Copas (1973) claims that the likelihood is always maximized at a joint distribution of potential outcomes that preserves the direct estimates of the marginal distributions of potential outcomes from equation (4) and which has either the maximal or minimal number of defiers. In practice, we often find that the likelihood is maximized at a distribution that does not preserve these marginal estimates. However, in all applications we have considered, the likelihood is maximized at either the estimated upper or lower Fréchet bound on defiers.

2.4.2 The Maximum Likelihood Decision Rule is Bayes Optimal

The maximum likelihood decision rule is Bayes optimal under appropriate specifications of the decision maker's utility function and subjective prior.[3] For the former, we specify a discrete utility function, which assigns zero utility for incorrect guesses and an arbitrary amount of strictly positive utility for correct guesses; for the latter, we specify a prior that is uniform over all possible distributions of potential outcomes for n subjects. Appendix D.2 reviews the formal definition and proof of Bayes optimality under these conditions.

While the conditions under which our decision rule is optimal are potentially restrictive, we emphasize that the design-based likelihood can be used to compute the Bayes optimal decision rule for any given loss function and prior, as we review in Appendix D. To justify our adoption of a uniform prior, we envision the sample as drawn IID from some population with an unknown joint distribution of potential outcomes (that is, the population shares of always takers, compilers, defiers, and never takers). We adopt a Dirichlet prior over the population distribution of potential outcomes, which is a common choice in sampling models, as it is the conjugate prior of the categorical and multinomial distributions. Our uniform prior over the sample distribution of potential outcomes is consistent with a "flat" Dirichlet prior, with all concentration parameters set to one. Sampling inspired priors like this do not,

[3]We thank Andriy Norets and Thomas Weimann for bringing the standard conditions that establish Bayes optimality to our attention.

in general, lead to flat posteriors within Fréchet sets. Regardless of the choice of prior, the ratio of the prior beliefs between any two distributions updates as long as the likelihood ratio differs from one. Adoption of a Bayesian perspective also allows us to construct credible sets for the joint distribution of potential outcomes. In Appendix E, we describe our algorithm for constructing smallest credible sets for the joint distribution of potential outcomes, as well as for the number of each type individually.

Bayes optimality ensures that our proposed decision rule does at least as well as all other decision rules under these optimality conditions. To quantify the gains from utilizing the optimal rule, we compare the Bayes expected utility obtained under the optimal rule to two alternative decision rules. The first alternative decision rule, which we refer to as the "Fréchet rule," estimates the marginal distributions of potential outcomes directly according to (4), and then chooses randomly among the joint distributions of potential outcomes in the corresponding Fréchet set. The Fréchet rule $\hat{\boldsymbol{\theta}}_F$ satisfies:

$$\mathbb{P}(\hat{\boldsymbol{\theta}}_F(\boldsymbol{x}) = \boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\mathcal{F}(\hat{\theta}_{1\bullet}(\boldsymbol{x}), \hat{\theta}_{\bullet 1}(\boldsymbol{x}); n)} & \text{if } \boldsymbol{\theta} \in \mathcal{F}(\hat{\theta}_{1\bullet}(\boldsymbol{x}), \hat{\theta}_{\bullet 1}(\boldsymbol{x}); n) \\ 0 & \text{otherwise.} \end{cases}$$

This decision rule captures the behavior of a statistician who assumes that the experimental data is uninformative within Fréchet sets and is therefore indifferent between any two distributions within the same Fréchet set.

The second alternative decision rule, which we refer to as the "monotonicity rule," chooses the maximizer of the design-based likelihood among those distributions with zero defiers and/or zero compilers: let $\widehat{\boldsymbol{\Theta}}_M(\boldsymbol{x})$ represent the set of joint distributions that maximize the likelihood among those distributions with zero compilers and/or defiers:

$$\widehat{\boldsymbol{\Theta}}_M(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \text{ s.t. } \theta_{01} = 0 \text{ or } \theta_{10} = 0.$$

Then, the monotonicity decision rule $\hat{\boldsymbol{\theta}}_M$ satisfies:

$$\mathbb{P}(\hat{\boldsymbol{\theta}}_M(\boldsymbol{x}) = \boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\widehat{\boldsymbol{\Theta}}_M(\boldsymbol{x})} & \text{if } \boldsymbol{\theta} \in \widehat{\boldsymbol{\Theta}}_M(\boldsymbol{x}) \\ 0 & \text{otherwise.} \end{cases}$$

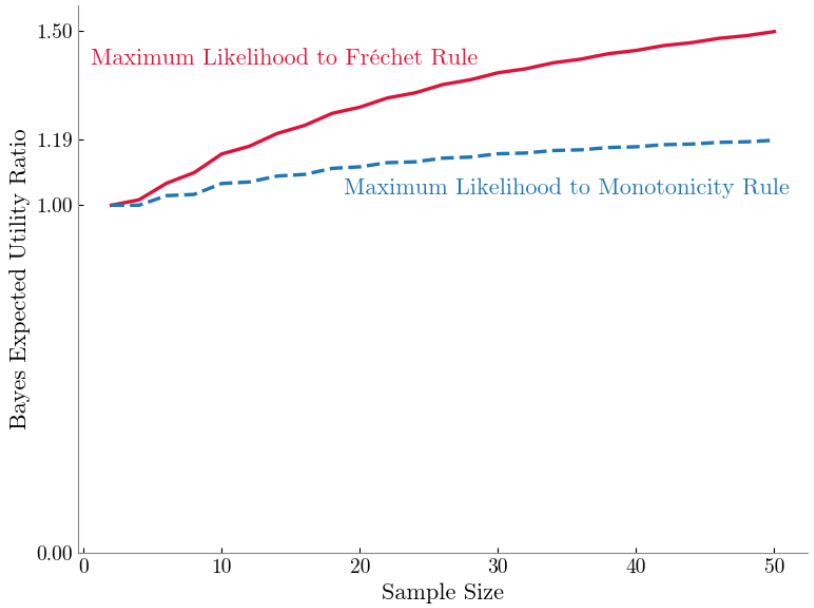
This decision rule approximates the behavior of a statistician who adopts a within-sample analogue to the monotonicity assumption of Imbens and Angrist (1994). This rule assumes that the statistician still knows the likelihood function, and is thus the "best" decision rule among those that adopt this monotonicity assumption.

We calculate the exact Bayes expected utility achieved by our optimal maximum likelihood rule and each of the two alternative rules. For a given sample size, we enumerate each of the finitely many feasible sample distributions of potential outcomes

$\boldsymbol{\theta}$. For each distribution $\boldsymbol{\theta}$, we compute the expected utility of our decision rules by comparing the distribution of estimates they produce under each data realization $\boldsymbol{x}$ to the true distribution. We average the resulting utilities over the values of $\boldsymbol{x}$ according to the likelihood. We then average these expected utilities according to a uniform prior for $\boldsymbol{\theta}$ to get the Bayes expected utility (see Appendix D for additional detail).

Figure 4 plots the ratio of the Bayes expected utility achieved by our proposed maximum likelihood rule to the Bayes expected utility achieved by each of the two alternative rules for even sample sizes between 2 and 50. The gains from using the optimal rule over either of the alternatives increase with the sample size. In a sample size of 50, the Bayes expected utility of the proposed maximum likelihood rule is 1.5 times higher than that of the Fréchet rule, and 1.19 times higher than that of the monotonicity rule.

Figure 4: Under Optimality Conditions, For Increasing Sample Sizes, Performance of Proposed Maximum Likelihood Rule Increases Relative to Fréchet and Monotonicity Rules



2.4.3 The Maximum Likelihood Decision Rule Maximizes Entropy

We can also motivate our maximum likelihood rule through the principle of maximum entropy (Jaynes, 1957a,b), which unites the theory of information with statistical physics. In Section 1.1, we discuss how differences in entropy drive variation in the likelihood such that the maximum likelihood distribution is also the maximum entropy distribution. Jaynes (1968) advocates for application of the principle of maximum entropy to circumvent the subjective specification of a prior. To make Bayesian decision-making more objective, one option is to choose the least informative prior.

Jaynes' alternative to make the process more objective is to choose the least informative updated distribution. The distribution that can generate the data in the greatest number of ways—the distribution that maximizes entropy, also the distribution that maximizes our design-based likelihood—is the least informative updated distribution. In Appendix F, we illustrate with a figure similar to Figure 1 that in the Monty Hall game, a celebrated application of Bayesian decision making, it is possible to make the same decision without specifying a Bayesian prior using a maximum likelihood decision rule and the principle of maximum entropy.

3 Applications: Counting Defiers

We apply our rule to two published experiments with positive, statistically significant average effects on takeup of desired health behaviors and plausible defiers. Figure 3 summarizes the results from all possible empirical applications in experiments with given sizes, illustrating empirical conditions under which our rule yields defiers. It is predictable from the data for each application and the patterns in Figure 3 that our first application yields no defiers, and our second application yields defiers, which we confirm with exhaustive grid search. Our focus here is to provide examples for applied researchers to follow when presenting and interpreting quantitative results and to provide substantive insights. The top of Table 1 summarizes the context and design for both applications.

In Tappin et al. (2015), the sample includes 612 pregnant smokers, and the definition of "takeup" is to quit smoking, which is a desired health behavior given concerns that smoking can cause pregnancy complications and lasting adverse effects on children. The experiment uses a completely randomized design, with half of the subjects assigned to an intervention arm with a payment and the other half assigned to a control arm with usual care. Monotonicity assumptions have a long history in the context of interventions that induce pregnant women to quit smoking. In a study discussed by Angrist, Imbens, and Rubin (1996), Permutt and Hebel (1989) proposed an early monotonicity assumption requiring that women who would quit smoking in control would also quit smoking in an arm that received a multi-faceted intervention. However, literature in psychology and economics expresses concern that payments can backfire in many contexts, including on-time pickup of children from day care (Gneezy and Rustichini, 2000), blood donation (Mellström and Johannesson, 2008; Lacetera and Macis, 2010), and vaccination (Schneider, Campos-Mercade, Meier, Pope, Wengström, and Meier, 2023). Defiers are plausible in Tappin et al. (2015) because a payment could backfire by crowding out pregnant women's intrinsic motivations to quit smoking based on moral ideals and health goals.

The average effect of the payment relative to usual care is a statistically significant 14 percentage point increase in the smoking quit rate on a base of 8% in control. Our proposed decision rule yields evidence beyond the average effect: our maximum likelihood estimate includes $52/612 = 8\%$ always takers (effect=0), $86/612 = 14\%$ compliers (effect=1), and $474/612 = 77\%$ never takers (effect=0). The weighted

Table 1: Applications: Standard Statistics and Proposed Design-Based Maximum Likelihood Estimates

	Tappin et al. (2015)	**Johnson and Goldstein (2003)**
Context and Design		
Takeup	Quit Smoking	Become Organ Donor
Intervention	Payment	Opt-out
Control	Usual Care	Opt-in
Sample	Pregnant Smokers	Online Respondents
Design	Completely Randomized	Bernoulli Randomized
Standard Statistics		
Average Effect	$69/306 - 26/306 = 14\%$	$50/61 - 23/54 = 39\%$
[95% Confidence Interval]	$[8\%, 20\%]$	$[23\%, 56\%]$
Fisher's Exact Test p-value	< 0.001	< 0.001
Intervention Takeup Rate	$69/306 = 23\%$	$50/61 = 82\%$
Control Takeup Rate	$26/306 = 8\%$	$23/54 = 43\%$
Sample Size	612	115
Proposed Design-Based Maximum Likelihood Estimates		
Always Takers	$52/612$=8%	$28/115$=24%
[95% Smallest Credible Set]	$[0/612, 72/612]$=[0%, 12%]	$[0/115, 63/115]$=[0%, 55%]
Compliers	$86/612$=14%	$66/115$=57%
[95% Smallest Credible Set]	$[44/612, 164/612]$=[7%, 27%]	$[23/115, 81/115]$=[20%, 70%]
Defiers	$0/612$=0%	$21/115$=18%
[95% Smallest Credible Set]	$[0/612, 71/612]$=[0%, 12%]	$[0/115, 34/115]$=[0%, 30%]
Never Takers	$474/612$=77%	$0/115$=0%
[95% Smallest Credible Set]	$[388/612, 498/612]$=[63%, 81%]	$[0/115, 32/115]$=[0%, 28%]

Percentages are rounded to whole numbers. As discussed in Appendix A, we construct a 95% smallest credible set on the number of each type in the sample by reporting the collection of values for each type occurring within the 95% smallest credible set on the joint distribution of potential outcomes. Because the numbers of always takers, compliers, defiers, and never takers must sum to the sample size, the 95% smallest credible sets for each of the four types are dependent.

average effect across the three types $(8 \times 0 + 14 \times 1 + 77 \times 0)$ is equal to the estimated average effect of 14 percentage points, which is also equal to the share of compliers. We would obtain the same results if we assumed monotonicity in the sample, so our rule provides evidence in favor of monotonicity without assuming it.

We emphasize that even though our proposed decision rule provides evidence that there are no defiers in the sample, there is considerable uncertainty in our maximum likelihood estimate. As discussed in Appendix E, we report a 95% smallest credible

set on the number of defiers in the sample, which we construct by reporting the collection of numbers of defiers occurring within the 95% smallest credible set on the joint distribution of potential outcomes in the sample. We caution that because the numbers of always takers, compliers, defiers, and never takers must sum to the sample size, the credible sets for each of the four types are dependent. The smallest credible set on defiers includes up to 71 defiers, which represents 71/612=12% of the sample size.

Building on the evidence provided by our decision rule, researchers could go on to make the stronger assumption of monotonicity and proceed as usual. With additional data on covariates, they could calculate average characteristics of always takers, compliers, and never takers (Imbens and Rubin, 1997; Katz, Kling, Liebman, et al., 2001; Abadie, 2002, 2003). With additional data on a second stage outcome like birth weight, they could interpret the resulting instrumental variable estimate as a local average treatment effect on compliers (Imbens and Angrist, 1994).

Our second example, Johnson and Goldstein (2003), is a high profile experiment in which a simple change from an opt-in default to an opt-out default increases organ donation almost twofold from 43% in control to 82% in intervention. Results from Johnson and Goldstein (2003) have inspired state policy changes to encourage life-saving organ donations to combat chronic organ shortages (Kessler and Roth, 2025). However, in a recent experiment by Kessler and Roth (2025), in which takeup represents actually becoming an organ donor instead of hypothetically becoming an organ donor, the average effect is close to zero. Especially given the hypothetical outcome in Johnson and Goldstein (2003), both compliers and defiers are plausible. Compliers may think of the default as an endorsement by the experimenter and thus follow it regardless of what it is. Defiers may think of a default as a threat to their autonomy and thus go against it regardless of what it is.

Given the plausibility of defiers in this context, a monotonicity assumption might not be reasonable, but there is no obvious alternative, so standard practice is to engage with bounds on defiers, which can be large. As shown at the bottom of Table A.1, in which we report optional auxiliary statistics, the smallest possible number of defiers in the sample is 0 because it is always possible for all people who take up to be always takers and all people who do not take up to be never takers, as in the null of Fisher's exact test. The largest possible number of defiers in the sample occurs if all $11(= 61-50)$ people who do not take up in intervention and all 23 people who take up in control are defiers, yielding a maximum defier share of $30\%(= (23+11)/115)$. The estimated Fréchet bounds are tighter because they impose that the estimated marginal distribution of potential outcomes in one arm yields the true marginal distribution of potential outcomes in the other arm and thus in the sample. To calculate the upper Fréchet bound from (3), we consider the bounds on the share of defiers implied by the takeup rates in intervention and control. The intervention takeup rate implies that the share of defiers in intervention and thus in the sample is at most $18\%(= 100\%-82\%)$, and the control takeup rate implies that the share of defiers in control and thus in

the sample is at most 43%, so the share in intervention binds, and the upper Fréchet bound is 18%.

The Fréchet bounds are estimated with error, so there is even greater uncertainty in the number of defiers. The maximum number of defiers in our 95% smallest credible set is 34, which represents 30% of the sample of 115. The 95% confidence interval on the population fraction of defiers from Imbens and Angrist (1994) also accounts for estimation of the Fréchet bounds, so we report it for comparison. It is slightly narrower, extending from 0% to 26%, but still includes the estimated Fréchet bounds. Accounting for estimation of the Fréchet bounds is empirically important. As demonstrated in Figure 2, if we do not account for estimation of the Fréchet bounds and instead construct a 95% smallest credible set on defiers within the estimated Fréchet set, variation in the likelihood allows us to exclude 8 and 9 defiers—counts in the middle of the Fréchet bounds translated into whole numbers from 0 to 21.

The maximum likelihood estimate indicates that the sample includes $21/115 = 18\%$ defiers, which is equal to the estimated upper Fréchet bound. The maximum likelihood estimate also includes $28/115 = 24\%$ always takers, $66/115 = 57\%$ compliers, and no never takers. The maximum likelihood estimate is the same as the estimate that would be obtained under an assumption of no never takers. Under an assumption of no defiers, the estimated average effect of 39% is entirely due to compliers, who represent 39% of the sample. The maximum likelihood estimate preserves the estimated average effect ($39\% = 57\% \times (1) + 18\% \times (-1)$), but it tells a very different story. Compliers representing 57% of the sample are affected in one direction and defiers representing 18% of the sample are affected in the other. The total share of people affected in the maximum likelihood estimate of $75\% = (57\% + 18\%)$ is almost two times as large as the number of people affected under monotonicity.

In this application, evidence from the design-based model suggests a weaker alternative assumption to monotonicity—an assumption of no never takers—that can subsequently be imposed. While the assumption of no never takers is still strong, by the principle of maximum entropy, it "assumes the least" about the joint distribution of potential outcomes (Jaynes, 1968). From the design-based likelihood in (2), the assumption of no never takers can generate the data in (28 choose 15) $\times$ (66 choose 35) $\times$ (21 choose 11) $= 8.5 \times 10^{31}$ randomized assignments representing 3.4% of the total randomized assignments ((115 choose 61) $= 2.5 \times 10^{33}$). The assumption of no defiers can generate the data in only (49 choose 26) $\times$ (45 choose 24) $\times$ (21 choose 11) $= 7.8 \times 10^{31}$ randomized assignments, representing 3.1% of the total randomized assignments. An ex post justification for the assumption of no never takers is that though experimental subjects might feel comfortable complying with or defying defaults imposed by the experimenter, they might feel uncomfortable revealing to the experimenter that they would not donate their organs under either default, especially since they do not know when they are asked a hypothetical question if a later question will ask them for a response under the other default.

If a second stage outcome such as actual organ donation is available, under an

assumption of no never takers, the instrumental variable estimate is no longer interpretable as an average treatment effect on compliers without further assumptions. As discussed by Angrist, Imbens, and Rubin (1996), under an additional assumption that average treatment effects are the same for compliers and defiers, the instrumental variable estimate is again interpretable as a LATE. Tchetgen Tchetgen (2024) proposes an alternative assumption to interpret the instrumental variable estimand in the presence of compliers and defiers as a "nudge average treatment effect."

The value of an assumption of no never takers becomes apparent if data are available on covariates. Under the assumption of no never takers, it is possible to label some specific people as compilers and others as defiers: all people in intervention who do not take up must be compliers, and all people who do not take up in control must be defiers. This reasoning allows us to make claims that the intervention had a causal effect on some specific people in the sample. Using terminology from Pearl (1999), under an assumption of no never takers, the intervention was a "necessary" cause of no takeup for all 11 people assigned intervention who do not take up, and was is a "sufficient" cause of no takeup for all 31 people assigned control who do not take up. Under a monotonicity assumption, it is not possible to directly label any specific people as compliers or defiers because people who take up in intervention can be compliers or always takers, and people who do not take up in control can be compliers and never takers. An assumption that allows us to label some specific people as compilers and others as defiers could be an important first step to learn their characteristics, improve targeting in future samples, and obtain a larger average effect.

4 Contributions to the Literature

Our primary insight is that a randomized experiment can yield evidence on an estimand more primitive than the average effect: the joint distribution of potential outcomes in the sample. The evidence can support a monotonicity assumption or inform a specific alternative assumption. We obtain evidence by using the structure of the implemented randomization process to construct a design-based model and its resulting likelihood function, and we evaluate this evidence in a statistical decision theory framework. Our work contributes to the literature on design-based econometrics, statistical decision theory, and causal inference without monotonicity.

In his 1935 book on the design of experiments, Fisher notes that the *design* of Charles Darwin's experiments was "greatly superior" to the "statistical methods available at the time" (Fisher, 1935). In their 2017 handbook chapter on the econometrics of experiments, Susan Athey and Guido Imbens write that they "recommend using statistical methods that are directly justified by randomization, in contrast to the more traditional sampling-based approach that is commonly used in econometrics" (Athey and Imbens, 2017). The sampling-based approach assumes that people in an experiment are sampled randomly from a population, and it attempts to learn about the population. The design-based approach, which we adopt, attempts to learn about

the sample of people in the experiment, which is useful since participation in experiments and clinical trials can be far from random (Alsan, Cawley, Doyle, and Skelley, 2025) and some samples do not come from a meaningful population (Abadie, Athey, Imbens, and Wooldridge, 2020).

The design-based model enables our advances to the frontier. Our ability to learn about defiers with a design-based likelihood may seem paradoxical because it is well known that the traditional sampling-based likelihood does not vary with the number of defiers within the Fréchet bounds determined by the estimated marginal distributions of potential outcomes in intervention and control. As we review in Appendix C, for a given pair of marginal distributions, the likelihood function in the traditional sampling-based model is flat over all copulas connecting them. A large literature has focused on specifying what we can learn from the Fréchet bounds in the traditional sampling-based model (Balke and Pearl 1997; Heckman, Smith, and Clements 1997; Manski 1997a; Tian and Pearl 2000; Zhang and Rubin 2003; Imbens and Manski 2004; Fan and Park 2010; Mullahy 2018; Ding and Miratrix 2019; Li and Pearl 2019; Bai, Huang, Moon, Shaikh, and Vytlacil 2024; Semenova 2024).

The design-based model combines a model of randomization within an experiment with causal models of potential outcomes from Neyman (1923), Welch (1937), Kempthorne (1952), Copas (1973), Rubin (1974, 1977), Greenland and Robins (1986), Holland (1986) and others. Copas (1973) derives the design-based likelihood of a completely randomized experiment. We generalize the design-based model to derive the likelihood from a generic randomization process, which we demonstrate with the likelihood for Bernoulli randomization. While Copas (1973) uses the design-based likelihood to test hypotheses about the average effect and the Fisher null hypothesis, our advance is to use the likelihood to learn about a more primitive object: the joint distribution of potential outcomes in the sample, including the number of defiers. In doing so, we connect Copas' design-based model of an experiment to the modern causal model, which reveals that experimental data contains information about the number of defiers in the sample, even if it does not contain information about their population analog. Likelihood-based analysis is uncommon in design-based models of experiments, partly due to its association with the sorts of restrictive "parametric" assumptions that experiments are constructed to avoid. One notable exception is Li and Ding (2016), who construct exact confidence intervals on the average effect using parametric distributions implied by the experiment design. By combining the potential outcomes framework with a parametric model of the randomization process itself, which is under the control of the designer of the experiment, the design-based model yields a likelihood function free of restrictive assumptions.

Our use of a decision-theoretic rule in the style of Wald (1949), rather than a classical hypothesis test, facilitates our progress. Classical hypothesis tests specify a status-quo null hypothesis and place the burden of evidence on the alternative hypothesis. This asymmetric treatment of Type I and Type II errors can imply an unrealistically conservative decision maker (Tetenov, 2012) and exclude statistical

evidence that, while weak, is still informative. To make use of all of the evidence present in our design-based likelihood, we focus on the statistical decision problem of choosing the correct joint distribution of potential outcomes in the sample, rather than on testing hypotheses about a prespecified distribution under prespecified size and power. We do not claim to provide a "test" of a monotonicity assumption. We provide a "decision" about the joint distribution of potential outcomes that can provide evidence for or against a monotonicity assumption.

We contribute to the integration of statistical decision theory into econometrics (Manski, 2004; Dehejia, 2005; Manski, 2007; Hirano, 2008; Hirano and Porter, 2009; Stoye, 2012; Kitagawa and Tetenov, 2018; Manski, 2018, 2019; Manski and Tetenov, 2021; Fernández, Montiel Olea, Qiu, Stoye, and Tinda, 2024). Economics is a natural discipline to embrace statistical decision theory because statistical decision theory has such a tight relationship with economic theory. Statistical decision theory is natural to apply in our setting because it does not require large sample assumptions or asymptotic approximations. Our use of design-based statistical decision theory departs from "asymptotic analysis of statistical decision rules in econometrics" as reviewed in Hirano and Porter (2020) in favor of the literature that focuses on statistical decision theory in finite sample settings (Canner, 1970; Manski and Tetenov, 2007; Schlag, 2007; Stoye, 2007, 2009; Tetenov, 2012). Our auxiliary construction of credible sets on the joint distribution of potential outcomes in Table A.1 provides a Bayesian alternative to classical asymptotic inference on Fréchet bounds (Manski, Sandefur, McLanahan, and Powers, 1992; Horowitz and Manski, 2000; Tamer, Chernozhukov, and Hong, 2004) and the number of defiers within the Fréchet bounds (Imbens and Manski, 2004).

Our proposed decision rule engages with justifications commonly employed in statistical decision theory. As we show in Figure 3, our decision varies systematically with the data. One challenge with statistical decision criteria such as minimax and minimax regret is that researchers have shown instances in which they yield rules that do not vary with the data (Schlag, 2003; Manski, 2004; Hirano and Porter, 2009; Stoye, 2009). In his influential introduction to statistical decision theory, Ferguson (1967) recognizes that because maximum likelihood rules vary with the data, they are "reasonable" because they are better than "just guessing." "However," he continues, "decision theory, as developed here, is devoted to the problem of finding optimal rules, so that we do not refer to maximum likelihood estimates [...] unless they turn out naturally to be optimal in some sense." He goes on to claim that "because of the basic theorem of decision theory," which he refers to as the complete class theorem, "in general, every really good decision rule is practically Bayes." In Section 2.4.2, we further justify our decision rule as Bayes optimal under the appropriate conditions. The Bayes optimality of our rule enables us to make important claims. Specifically, it implies that our rule is admissible—that is, there is no alternative rule that performs at least as well for all possible joint distributions of potential outcomes, and strictly better for some (Ferguson, 1967). Furthermore, it ensures that our rule cannot be

bested in a betting framework (Freedman and Purves, 1969). It also allows us to quantify the relative losses from suboptimal rules, including a rule that places equal probability on all numbers of defiers within the Fréchet bounds, and a rule that assumes monotonicity, as we do in Section 2.4.2.

In addition to standard decision theoretic justifications, we also offer a perspective that our design-based maximum likelihood decision rule is optimal through the principle of maximum entropy (Jaynes, 1957a,b), which does not require specification of a prior. Golan (2002) reviews a literature in econometrics that uses the principle of maximum entropy to abstract away from functional form assumptions on the likelihood function. In our setting, the design of the randomization, which is under the control of the experimenter, informs the functional form of the likelihood function, which also determines the functional form of the entropy.

Our results inform monotonicity assumptions, which have become widespread. Influential work by Imbens and Angrist (1994) proposes a monotonicity assumption that assumes away compilers or defiers in the first stage of an instrumental variable model, and Manski (1997b) proposes an analogous assumption in the reduced form. Monotonicity assumptions are ubiquitous in the analysis of first stages and they are gaining traction in the analysis of reduced forms (Alsan, Cawley, Doyle, and Skelley, 2025). First stage monotonicity assumptions are useful for the interpretation of instrumental variable estimates as local average treatment effects on compliers, but they can be difficult to defend in some contexts, particularly in health care settings where interventions may be helpful for some but harmful for others. We interpret our binary outcome as "takeup" in our applications so that we can use the familiar terminology of always and never "takers" developed for the first stage (Angrist, Imbens, and Rubin, 1996), but there is nothing in our approach that requires our binary outcome to be interpretable as "takeup." There is also nothing in our approach that requires our binary outcome to be interpretable as a "treatment" on a second stage outcome. Given concerns about side effects in medicine, important applications could involve a binary outcome that represents death.

Previous approaches have revealed evidence about the joint distribution of potential outcomes that violates monotonicity, but only with additional data beyond our binary intervention and outcome. Machine learning approaches of Wager and Athey (2018) and Semenova (2024) require rich data on covariates to reveal heterogeneous effects. Analysis of side effects in medicine requires data on secondary outcomes (Bernard et al., 2001). Specification tests of instrumental variable model assumptions can reveal evidence against monotonicity in the first stage (Imbens and Rubin, 1997; Richardson and Robins, 2010; Huber and Mellace, 2012, 2015; Kitagawa, 2015; Mourifié and Wan, 2017; Machado, Shaikh, and Vytlacil, 2019), and marginal treatment effect methods can reveal evidence against monotonicity in the second stage (Björklund and Moffitt, 1987; Heckman and Vytlacil, 1999; Kowalski, 2023a,b), but such approaches require the additional structure of a two-stage model as well as data on both stages. Evidence against first stage monotonicity by Chan, Gentzkow, and

Yu (2022) requires a multi-valued instrument and the ability to observe a second stage outcome for one value of the first stage outcome. We contribute to the literature on violations of monotonicity by demonstrating that, even without additional data, our design-based model of an experiment can reveal evidence about monotonicity.

We view the development of our visualizations as a secondary contribution. Our visualization of potential outcomes in Figure 1 allows us to depict how the same data can arise from different distributions of potential outcomes within a Fréchet set. Our visualization of how the MLE varies with the data in Figure 3 allows us to observe patterns that would be difficult to infer from isolated empirical applications.

We also offer substantive insights in actual empirical applications. In one, a payment intervention to induce pregnant women to quit smoking, we find no defiers. In another, a nudge intervention to get more people to hypothetically donate their organs, we find both compilers and the upper Fréchet bounds on defiers. In both cases, the weak evidence that we find can motivate stronger assumptions that researchers can use to learn more.

5 Implications for Future Econometric and Applied Research

We see our demonstration that a design-based model can yield evidence beyond the average effect with very minimal assumptions and data as a proof-of-concept that could pave the way for the production of stronger evidence that will achieve the ultimate goal of personalizing decisions. Even if the average effect is the decision-relevant parameter in some contexts, learning about heterogeneity in the average effect is a first step toward targeting that can ultimately yield a larger average effect. In our second empirical application, under the maximum likelihood joint distribution of potential outcomes, which includes no never takers, we can deduce that the people who are untreated in intervention must be defiers, and the people who are untreated in control must be compliers. Under monotonicity, it is never possible to deduce that a given individual is a complier, and defiers are assumed away. Our hope is that the ability to look directly at some compliers and defiers under alternative assumptions motivated by our statistical decision rule will pave the way to characterize them, target interventions toward compilers and away from defiers, and improve the average effect of future interventions.

Our work suggests several areas for further econometric research. Design-based likelihoods can be derived for more complex experimental settings, including those with covariates, two-stage models, and models with alternative randomization approaches, such as matched pair, stratified, and permuted block designs. These design-based likelihoods can be used to determine exact distributions of experimental data, test statistics, or confidence intervals without resorting to simulation methods, as is common in the randomization inference literature. Approximation methods and advances in algorithmic efficiency would facilitate these calculations in large experiments. The design-based likelihood from an experiment can be integrated into a variety of choice problems beyond the specific statistical decision problem considered

here. Optimal estimators under alternative utility functions or priors can be derived. Or, a design-based likelihood can be incorporated into treatment choice problems that seek to maximize average outcomes or novel targets including outcome quantiles (Guggenberger, Mehta, and Pavlov, 2024), welfare (Cui and Han, 2023), or counterfactual expected utilities (Ben-Michael, Imai, and Jiang, 2024; Christy and Kowalski, 2024b). Results from these new lines of inquiry could be used to engage with previous results on optimal experimental design that focus on estimating the average effect (Bai, 2022).

Design-based methods have been around for a long time, and we have demonstrated that it is feasible to implement our design-based decision rule, but a practical impediment to widespread implementation is that applied researchers do not report enough information about their experiments to replicate them with design-based methods (Young, 2019). We use large language models to pull (openAI's GPT-4o-mini) and categorize (openAI's GPT-o3-mini) the randomization processes from 2,080 papers associated with randomized controlled trials from the Abdul Latif Jameel Poverty Action Lab (J-PAL). Only 61% have enough description for the large language model to confidently categorize the randomization method from the paper or associated AEA RCT Registry entry. However, of those, 78% use designs captured by design-based likelihoods that we consider. In the full set of papers, over 60% describe some mechanism through which there could be defiers. Another practical impediment to design-based methods is that instead of reporting absolute numbers, several researchers only report rates or regression results. Mandated reporting of experimental design and standard sample counts by J-PAL, medical journals, and other authorities could facilitate the application of our design-based decision rule, allowing researchers to learn more from their experiments.

Appendix

Appendix A Optional Auxiliary Statistics

Figure A.1: Applications: Standard Statistics, Design-Based Maximum Likelihood Estimates, and Optional Statistics

	Tappin et al. (2015)	**Johnson and Goldstein (2003)**
Context and Design		
Takeup	Quit Smoking	Become Organ Donor
Intervention	Payment	Opt-out
Control	Usual Care	Opt-in
Sample	Pregnant Smokers	Online Respondents
Design	Completely Randomized	Bernoulli Randomized
Standard Statistics		
Average Effect	$69/306 - 26/306 = 14\%$	$50/61 - 23/54 = 39\%$
[95% Confidence Interval]	$[8\%, 20\%]$	$[23\%, 56\%]$
Fisher's Exact Test p-value	< 0.001	< 0.001
Intervention Takeup Rate	$69/306 = 23\%$	$50/61 = 82\%$
Control Takeup Rate	$26/306 = 8\%$	$23/54 = 43\%$
Sample Size	612	115
Proposed Design-Based Maximum Likelihood Estimates and Auxiliary Statistics		
Always Takers	52/612=8%	28/115=24%
[95% Smallest Credible Set]	$[0/612, 72/612]$=[0%, 12%]	$[0/115, 63/115]$=[0%, 55%]
[Largest Possible Support]	$[0/612, 95/612]$=[0%, 16%]	$[0/115, 73/115]$=[0%, 63%]
[Estimated Fréchet Bounds]	$[0/612, 52/612]$=[0%, 8%]	$[28/115, 49/115]$=[24%, 43%]
[Imbens Manski 95% Confidence Interval]	[0%, 11%]	[11%, 54%]
[95% Smallest Credible Set within Estimated Fréchet Bounds]	$[0/612, 10/612] \cup [14/612, 52/612]$=[0%, 2%] ∪ [2%, 8%]	$[28/115, 39/115] \cup [42/115, 49/115]$=[24%, 34%] ∪ [37%, 43%]
Compliers	86/612=14%	66/115=57%
[95% Smallest Credible Set]	$[44/612, 164/612]$=[7%, 27%]	$[23/115, 81/115]$=[20%, 70%]
[Largest Possible Support]	$[0/612, 349/612]$=[0%, 57%]	$[0/115, 81/115]$=[0%, 70%]
[Estimated Fréchet Bounds]	$[86/612, 138/612]$=[14%, 23%]	$[45/115, 66/115]$=[39%, 57%]
[Imbens Manski 95% Confidence Interval]	[9%, 26%]	[26%, 68%]
[95% Smallest Credible Set within Estimated Fréchet Bounds]	$[86/612, 124/612] \cup [128/612, 138/612]$=[14%, 20%] ∪ [21%, 23%]	$[45/115, 52/115] \cup [55/115, 66/115]$=[39%, 45%] ∪ [48%, 57%]
Defiers	0/612=0%	21/115=18%
[95% Smallest Credible Set]	$[0/612, 71/612]$=[0%, 12%]	$[0/115, 34/115]$=[0%, 30%]
[Largest Possible Support]	$[0/612, 263/612]$=[0%, 43%]	$[0/115, 34/115]$=[0%, 30%]
[Estimated Fréchet Bounds]	$[0/612, 52/612]$=[0%, 8%]	$[0/115, 21/115]$=[0%, 18%]
[Imbens Manski 95% Confidence Interval]	[0%, 11%]	[0%, 26%]
[95% Smallest Credible Set within Estimated Fréchet Bounds]	$[0/612, 38/612] \cup [42/612, 52/612]$=[0%, 6%] ∪ [7%, 8%]	$[0/115, 7/115] \cup [10/115, 21/115]$=[0%, 6%] ∪ [9%, 18%]
Never Takers	474/612=77%	0/115=0%
[95% Smallest Credible Set]	$[388/612, 498/612]$=[63%, 81%]	$[0/115, 32/115]$=[0%, 28%]
[Largest Possible Support]	$[0/612, 517/612]$=[0%, 84%]	$[0/115, 42/115]$=[0%, 37%]
[Estimated Fréchet Bounds	$[422/612, 474/612]$=[69%, 77%]	$[0/115, 21/115]$=[0%, 18%]
[Imbens Manski 95% Confidence Interval]	[64%, 81%]	[0%, 26%]
[95% Smallest Credible Set within Estimated Fréchet Bounds]	$[422/612, 432/612] \cup [436/612, 474/612]$=[69%, 71%] ∪ [71%, 77%]	$[0/115, 11/115] \cup [14/115, 21/115]$=[0%, 10%] ∪ [12%, 18%]

Percentages are rounded to whole numbers. As discussed in Appendix A, we construct a 95% smallest credible set on the number of each type in the sample by reporting the collection of values for each type occurring within the 95% smallest credible set on the joint distribution of potential outcomes. Because the numbers of always takers, compliers, defiers, and never takers must sum to the sample size, the 95% smallest credible sets for each of the four types are dependent.

Appendix B Derivation of the Design-Based Likelihood

The design-based likelihood of the joint distribution of potential outcomes in the sample $\boldsymbol{\theta} = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ given the data $\boldsymbol{x}$ is the probability of this realization of the data given $\boldsymbol{\theta}$. To derive the distribution of the data, first let us define the joint distribution of potential outcomes in the intervention arm $\boldsymbol{I} \equiv \left(I_{11}, I_{10}, I_{01}, I_{00}\right)$ as a random vector whose elements represent the numbers of always takers, compliers, defiers, and never takers randomized into intervention. These values are unobserved to the experimenter. However, the randomization process—the design of the experiment—implies a data generating process for the distribution of potential outcomes in the intervention arm $\boldsymbol{I}$ conditional on the distribution of potential outcomes in the sample $\boldsymbol{\theta}$, which we can then use to derive the distribution of the data $\boldsymbol{X}$.

Consider an experiment where the randomization design assigns subjects to the intervention or control arm through a series of Bernoulli trials. Then, each subject's probability of assignment to intervention is p, and assignments are independent across

subjects as well as across principal strata. This independence allows us to write the distribution of $\boldsymbol{I}$ as the product of four independent binomial distributions:

$$\mathbb{P}\Big(I_{11} = I_{11}, I_{10} = I_{10}, I_{01} = I_{01}, I_{00} = I_{00} \mid \boldsymbol{\theta}\Big) = \binom{\theta_{11}}{I_{11}}\binom{\theta_{10}}{I_{10}}\binom{\theta_{01}}{I_{01}}\binom{\theta_{00}}{I_{00}} \times p^{\sum_{j,k} i_{j,k}}(1-p)^{n-\sum_{j,k} i_{j,k}}. \quad (5)$$

Alternatively, in a completely randomized experiment, the randomization design fixes the number of subjects in the intervention arm m and selects any of the possible combinations of m subjects in intervention and $n-m$ subjects in control with equal probability, as though drawing names from a hat. Under this design, $\boldsymbol{I}$ follows a multivariate hypergeometric distribution:

$$\mathbb{P}\Big(I_{11} = I_{11}, I_{10} = I_{10}, I_{01} = I_{01}, I_{00} = I_{00} \mid \boldsymbol{\theta}\Big) = \frac{\binom{\theta_{11}}{I_{11}}\binom{\theta_{10}}{I_{10}}\binom{\theta_{01}}{I_{01}}\binom{\theta_{00}}{I_{00}}}{\binom{n}{m}}. \quad (6)$$

In either case, the experimental design implies a data-generating process for the number of subjects of each type randomized into the intervention arm. While the distribution of potential outcomes in the intervention arm $\boldsymbol{I}$ is unobservable, we can use its data generating process to derive the distribution of the observable data $\boldsymbol{X}$. Note that each subject randomized into the intervention arm with outcome $Y = 1$ must be either an always taker or a complier: $X_{I1} = I_{11} + I_{10}$. Each subject randomized into the intervention arm with outcome $Y = 0$ must be either a never taker or a defier: $X_{I0} = I_{00} + I_{01}$. In the control arm, those observed with outcome $Y = 1$ must be either always takers that were not randomized into intervention, or defiers that were not randomized into intervention: $X_{C1} = \theta_{11} - I_{11} + \theta_{01} - I_{01}$. And finally, those in the control arm with outcome $Y = 0$ must be either never takers that were not randomized into intervention, or compliers that were not randomized into intervention: $X_{C0} = \theta_{00} - I_{00} + \theta_{10} - I_{10}$. Thus, we can write the probability of the observed data $\boldsymbol{X}$ conditional on the joint distribution of potential outcomes $\boldsymbol{\theta}$ as:

$$\begin{aligned}\mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\big) &= \mathbb{P}\Big(I_{11} + I_{10} = x_{I1},\ \big(\theta_{11} - I_{11}\big) + \big(\theta_{01} - I_{01}\big) = x_{C1}, \\ &\qquad I_{00} + I_{01} = x_{I0},\ \big(\theta_{00} - I_{00}\big) + \big(\theta_{10} - I_{10}\big) = x_{C0} \mid \boldsymbol{\theta}\Big) \\ &= \mathbb{P}\Big(I_{11} + I_{10} = x_{I1},\ I_{11} + I_{01} = x_{C1} - \theta_{11} - \theta_{01}, \\ &\qquad I_{00} + I_{01} = x_{I0},\ I_{00} + I_{10} = x_{C0} - \theta_{00} - \theta_{10} \mid \boldsymbol{\theta}\Big).\end{aligned}$$

A realization of the data $\boldsymbol{X}$ may be produced from more than one realization of the distribution of potential outcomes in intervention $\boldsymbol{I}$. To find the probability of a realization of $\boldsymbol{X}$, we sum together the probabilities of each realization of $\boldsymbol{I}$ that could have produced it. We can index these realizations through the realization i of I_{11} and

solve the following system of equations for the elements of $\boldsymbol{I}$:

$$\begin{aligned} I_{11} + I_{10} &= x_{I1}, \\ I_{11} + I_{01} &= \theta_{11} + \theta_{01} - x_{C1}, \\ I_{11} + I_{10} + I_{01} + I_{00} &= x_{I1} + x_{I0}, \\ I_{11} &= i. \end{aligned}$$

Rearranging yields

$$\begin{aligned} I_{11} &= i \\ I_{10} &= x_{I1} - i \\ I_{01} &= \theta_{11} + \theta_{01} - x_{C1} - i \\ I_{00} &= x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}. \end{aligned}$$

The value i is restricted to the set $\mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})$ such that $\boldsymbol{I}$ remains within the support implied by $\boldsymbol{\theta}$, namely $0 \leq I_{11} \leq \theta_{11}$, $0 \leq I_{10} \leq \theta_{10}$, $0 \leq I_{01} \leq \theta_{01}$, and $0 \leq I_{00} \leq \theta_{00}$. The probability of a realization of $\boldsymbol{X}$ is just the sum of the probability of each of these realizations of $\boldsymbol{I}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\big) = \sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} \mathbb{P}\Big(& I_{11} = i, \\ & I_{10} = x_{I1} - i, \\ & I_{01} = \theta_{11} + \theta_{01} - x_{C1} - i, \\ & I_{00} = x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01} \mid \boldsymbol{\theta}\Big). \end{aligned} \tag{7}$$

Equation (7) represents the general design-based likelihood, where the probability of each distribution of potential outcomes in intervention $\boldsymbol{I}$ given the distribution of potential outcomes in the sample $\boldsymbol{\theta}$ is informed by the design of the experiment. When subjects are assigned to intervention by a series of Bernoulli trials, $\boldsymbol{I}$ follows

the distribution in (5), yielding:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \\ & \times p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}}. \end{aligned} \tag{8}$$

Alternatively, in a completely randomized experiment, $\boldsymbol{I}$ follows the distribution in (6), yielding:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{m + x_{C1} + i - \theta_{11} - \theta_{01} - x_{I1}} \Big/ \binom{n}{m} \end{aligned} \tag{9}$$

where we have substituted $m = x_{I1} + x_{I0}$. In either case, once conditioning on the realized data $\boldsymbol{x}$, the likelihood is proportional to

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{m + x_{C1} + i - \theta_{11} - \theta_{01} - x_{I1}} \end{aligned} \tag{10}$$

This final expression is equivalent to the likelihood in Copas (1973).

Appendix C Derivation of a Canonical Sampling-Based Likelihood

Suppose that subjects are sampled into the experiment with replacement from a population: let q_{11} represent the share of the population who are always takers q_{10} the share of compliers, q_{01} the share of defiers , and q_{00} the share never takers. Sampling with replacement implies that each subject's vector of potential outcomes is independently and identically distributed (IID) according to these shares.

We can derive likelihood functions for the population distribution of potential outcomes $\boldsymbol{q} \equiv (q_{11}, q_{10}, q_{01}, q_{00})$ given the data from the experiment, depending on the randomization procedure into intervention. If subjects are assigned to intervention or control via a series of Bernoulli trials, we can view the combination of assignment Z and outcome Y across the n subjects as independent categorical variables whose distribution is defined by the population distribution of potential outcomes $\boldsymbol{q}$ and the probability of assignment to intervention p:

$$\begin{aligned}
\mathbb{P}\big(Y = 1, Z = I\big) &= (q_{11} + q_{10})p \\
\mathbb{P}\big(Y = 0, Z = I\big) &= \big(1 - (q_{11} + q_{10})\big)p \\
\mathbb{P}\big(Y = 1, Z = C\big) &= (q_{11} + q_{01})(1 - p) \\
\mathbb{P}\big(Y = 0, Z = C\big) &= \big(1 - (q_{11} + q_{01})\big)(1 - p).
\end{aligned}$$

The experimental data $\boldsymbol{X}$ are counts of realizations of independent categorical variables. Therefore, $\boldsymbol{X}$ follows the multinomial distribution with the probabilities above, which yields the likelihood expression:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) &= \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{q}) \\
&= \frac{n!}{x_{I1}!x_{I0}!x_{C1}!x_{C0}!} p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \\
&\quad \big(q_{11} + q_{10}\big)^{x_{I1}} \big(1 - (q_{11} + q_{10})\big)^{x_{I0}} \\
&\quad \big(q_{11} + q_{01}\big)^{x_{C1}} \big(1 - (q_{11} + q_{01})\big)^{x_{C0}}. \qquad (11)
\end{aligned}$$

This likelihood is equivalent to a likelihood appearing in Barnard (1947).

Alternatively, if the experiment is "completely randomized," such that $m \leq n$ subjects are assigned to intervention by drawing names from a hat and the remainder are assigned to control, we can view the sample as two sets of Bernoulli trials. The first set of m trials represent the subjects sampled into the intervention arm, where there are X_{I1} treated subjects with independent probabilities of being treated equal to the probability that $Y_I = 1$ in the population, $q_{11} + q_{10}$. The second set of $n - m$ trials represent the subjects sampled into the control arm, where there are X_{C1} treated subjects with independent probabilities of being treated equal to the probability that $Y_C = 1$ in the population, $q_{11} + q_{01}$. The two sets of Bernoulli trials are independent, allowing us to write the likelihood of a completely randomized experiment as the

product of two Binomial random variables X_{I1} and X_{C1}:

$$\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \binom{m}{x_{I1}} \big(q_{11} + q_{10}\big)^{x_{I1}} \big(1 - (q_{11} + q_{10})\big)^{m - x_{I1}} \binom{n-m}{x_{C1}} \big(q_{11} + q_{01}\big)^{x_{C1}} \big(1 - (q_{11} + q_{01})\big)^{n - m - x_{C1}}. \tag{12}$$

This likelihood is equivalent to likelihoods from Barnard (1947) and Kline and Walters (2020).

In each of these likelihood functions, the joint distribution of potential outcomes in the population $\boldsymbol{q}$ appears only through the transformations $q_{11} + q_{10}$ and $q_{11} + q_{01}$, which are the marginal distributions of the potential outcomes in intervention and control in the population, respectively. Therefore, the likelihoods of any two distributions in the same Fréchet set are equal. That is, the experimental data contains no information about the joint distribution of potential outcomes in the population beyond the marginal distributions of potential outcomes.

Appendix D Bayes Optimality of the Maximum Likelihood Decision Rule

Appendix D.1 Preliminaries on Statistical Decision Theory

Suppose a decision maker wishes to guess the joint distribution of potential outcomes in the sample. We write the decision maker's guess as $\widehat{\boldsymbol{\theta}}$, and we define a utility function over a guess $\widehat{\boldsymbol{\theta}}$ and the true distribution $\boldsymbol{\theta}$. We focus on the following utility specification, which yields one util when the guess is correct and zero utils when the guess is incorrect:

$$u(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbf{1}_{\{\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}\}}.$$

We also allow the decision maker to choose a randomized guess, which ascribes a probability distribution over the possible values of $\boldsymbol{\theta}$. We define the decision maker's utility over a randomized guess ρ as the expected utility of guessing according to the probabilities ascribed by ρ:

$$U(\rho, \boldsymbol{\theta}) = \sum_{\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} u(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \rho(\widehat{\boldsymbol{\theta}}). \tag{13}$$

The decision maker chooses a decision rule that maps the observable data into (possibly) randomized guesses.[4] We write such a rule as $f : \boldsymbol{\mathcal{X}} \to \Delta(\boldsymbol{\Theta})$, where $\boldsymbol{\mathcal{X}}$ is the space of possible data realizations, $\boldsymbol{\Theta}$ is the space of possible distributions of

[4] We conflate here the standard definitions of "randomized decision rules" and "behavioral decision rules" (Ferguson, 1967) for expositional clarity. In settings of perfect recall, such as the setting we study here, the space of randomized and behavioral decision rules is equivalent (Kuhn, 1953).

potential outcomes, and $\Delta(\boldsymbol{\Theta})$ is the space of distributions over $\boldsymbol{\Theta}$. Given a true distribution of potential outcomes $\boldsymbol{\theta}$, the decision maker's expected utility from following a decision rule f is the expected value of $U(f(\boldsymbol{X}), \boldsymbol{\theta})$ with respect to the experimental outcome $\boldsymbol{X}$:

$$\begin{aligned} EU(f, \boldsymbol{\theta}) &= \mathbb{E}\big[U\big(f(\boldsymbol{X}), \boldsymbol{\theta}\big) \mid \boldsymbol{\theta}\big] \\ &= \sum_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}} \sum_{\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} u(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mathcal{L}(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{x}) f(\boldsymbol{x})(\widehat{\boldsymbol{\theta}}) \\ &= \sum_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) f(\boldsymbol{x})(\boldsymbol{\theta}). \end{aligned}$$

Under the specified utility function, the decision maker's expected utility is equal to the probability of guessing correctly.

One sense in which a decision rule can be optimal is in terms of Bayes expected utility: the average expected utility obtained according to some subjective prior distribution over $\boldsymbol{\theta}$:

$$EU_\pi(f) = \mathbb{E}_\pi\big[EU(f, \boldsymbol{\theta})\big] = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} EU(f, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}),$$

where $\pi \in \Delta(\boldsymbol{\Theta})$ is a subjective prior belief about $\boldsymbol{\theta}$. Decision rules that maximize this criterion are said to be "Bayes optimal."

Appendix D.2 Conditions for Bayes Optimality

In this appendix, we review that the result that the maximum likelihood decision rule is Bayes optimal when utility takes the form in (13) and the decision maker's subjective prior is uniform across realizations of the distribution of potential outcomes in the sample $\boldsymbol{\theta}$. To show this result, we first establish that the maximum *a posteriori* decision rule is Bayes optimal for our chosen loss function given any prior. The maximum *a posteriori* decision rule f^*_π selects the maxima of the posterior distribution of $\boldsymbol{\theta}$. Formally, let $\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})$ be the set of $\boldsymbol{\theta}$ values that maximize the posterior distribution given the observed data $\boldsymbol{x}$, i.e.

$$\begin{aligned} \widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x}) &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{P}\big(\boldsymbol{\theta} \mid \boldsymbol{x}\big) \\ &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \pi(\boldsymbol{\theta}), \end{aligned} \tag{14}$$

where $\pi \in \Delta(\boldsymbol{\Theta})$ is the prior belief about $\boldsymbol{\theta}$. The decision rule f^*_π can then be defined as follows:

$$f^*_\pi(\boldsymbol{x})(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}} & \text{if } \boldsymbol{\theta} \in \widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x}), \\ 0 & \text{o.w.} \end{cases} \tag{15}$$

Now, let g be an arbitrary decision function. The Bayes expected utility for decision function g is

$$\mathbb{E}\big[EU(g,\boldsymbol{\theta})\big] = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} EU(g,\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{16}$$

$$= \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) g(\boldsymbol{x})(\boldsymbol{\theta})\right] \pi(\boldsymbol{\theta}) \tag{17}$$

By rearranging terms in the summation, we can bound the Bayes expected utility of g:

$$\begin{aligned}
\mathbb{E}\big[EU(g,\boldsymbol{\theta})\big] &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) g(\boldsymbol{x})(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\Big) \\
&\leq \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(g(\boldsymbol{x})(\boldsymbol{\theta}) \max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \Big\{\mathcal{L}(\boldsymbol{\theta}' \mid \boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\}\Big) \\
&= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}} \left[\max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \Big\{\mathcal{L}(\boldsymbol{\theta}' \mid \boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\} \underbrace{\Big(\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g(\boldsymbol{x})(\boldsymbol{\theta})\Big)}_{=1}\right].
\end{aligned}$$

This bound is precisely the Bayes expected utility achieved by decision rule f^*_π:

$$\begin{aligned}
\mathbb{E}\big[EU(f^*_\pi,\boldsymbol{\theta})\big] &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\Theta} \Big(\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) f^*_\pi(\boldsymbol{x})(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\Big) \\
&= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})} \left(\frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}} \max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \Big\{\mathcal{L}(\boldsymbol{\theta}' \mid \boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\}\right) \\
&= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}} \left[\max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \Big\{\mathcal{L}(\boldsymbol{\theta}' \mid \boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\} \underbrace{\Big(\sum_{\boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}}\Big)}_{=1}\right]
\end{aligned}$$

Thus, since f^*_π achieves the upper bound on the Bayes expected utility of any decision rule, we conclude that f^*_π is Bayes optimal. Finally, observe that when the prior distribution $\pi(\theta)$ is constant, the maximizers of the posterior distribution in (14) are simply the maximizers of the likelihood, and the Bayes rule f^*_π in (15) simplifies to our maximum likelihood decision rule.

For alternative utility functions or priors, the optimal rule may not take a convenient form such as the MLE, but can still be computed by solving the linear programming problem of maximizing Bayes expected utility in (17) with respect to the (finitely-many) values of $g(\boldsymbol{x})(\boldsymbol{\theta})$, such that $\sum_{\boldsymbol{\theta}} g(\boldsymbol{x})(\boldsymbol{\theta}) = 1$ for each $\boldsymbol{x}$.

Appendix E Construction of Credible Sets

We construct credible sets on the joint distribution of potential outcomes in the sample using the posterior belief implied by a uniform prior and the design-based likelihood. A credible set of level α is a set of joint distributions of potential outcomes covering a posterior probability mass of at least α. We report "smallest credible sets", which is the credible set containing the fewest points. We construct these sets with the following algorithm:

1. Compute the posterior probability of every joint distribution of potential outcomes $\boldsymbol{\theta}$ given the observed data $\boldsymbol{x}$.
2. Order the joint distributions from highest to lowest posterior probability.
3. Compute the cumulative probability mass according to this ordering.
4. Starting with the highest probability distribution (which is the MLE), add joint distributions of potential outcomes to the credible set until the probability mass of the credible set exceeds α. In the case of ties, we add all distributions with the same probability simultaneously.

To report credible sets on the number of always takers, compilers, defiers, or never takers individually, we report the range of values for that type present in the credible set constructed on the joint distribution. Note that the credible sets on individual response types are not independent. For example, the maximal values in the individual credible sets for always takers, compilers, defiers, and never takers may not constitute an element of the credible set on the joint distribution of potential outcomes.

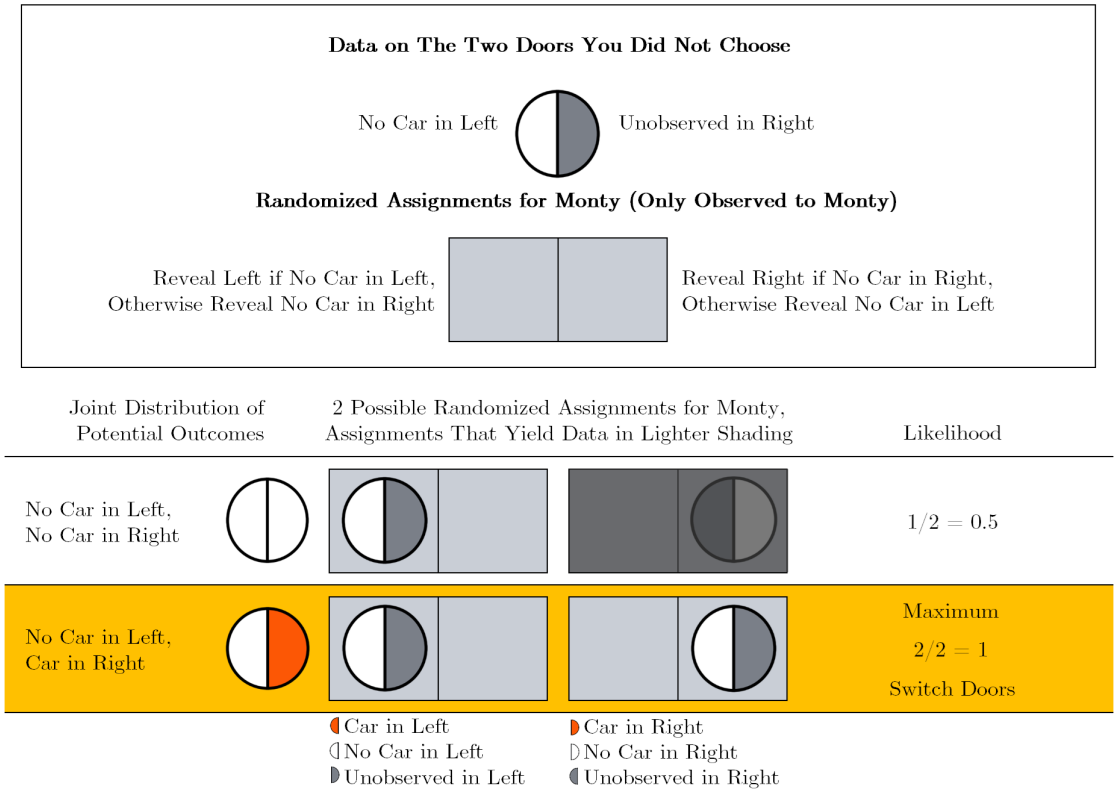
Appendix F Monty Hall: A Design-Based Model of the Monty Hall Game Can Reveal Evidence to Switch Doors

Suppose you are a contestant on the game show, "Let's Make a Deal," with host Monty Hall. He shows you three doors, he tells you that one has a car behind it, and he asks you to choose one. You get the car if you choose the correct door and nothing otherwise. You choose a door. Of the two doors you did not choose, Monty must open one door with no car behind it. Assume that a fair randomizer tells him which remaining door to open if he has two options. Without loss of generality, Monty reveals no car in the left door. Should you keep your door or switch doors?

By well-trodden Bayesian arguments, you should switch doors. Using our proposed visualization of potential outcomes, we illustrate here that the principle of maximum entropy can be used to arrive at the same decision without specifying a Bayesian prior. Wang, Tran, and Abebe (2016) discuss the principle of maximum entropy in the same context. We illustrate it here with our proposed visualization of potential outcomes.

Figure F.2 illustrates the Monty Hall problem using the approach to visualize potential outcomes that we develop to illustrate the design-based likelihood in Figure 1. In Figure 1, the left half of each circle represents the potential outcome in intervention, and the right half of each circle represents the potential outcome in control. The potential outcomes can take on values of "takeup" (orange), "no takeup" (white), or "unobserved" (grey). Here, the two halves of each circle represent the potential outcomes in the two doors that you did not choose: the left half represents the potential outcome in the left remaining door, and the right half represents the potential outcome in the right remaining door. The potential outcomes can take on values of "car" (orange), "no car" (white), or "unobserved" (grey).

Figure F.2: A Design-Based Model of the Monty Hall Game Can Reveal Evidence to Switch Doors



In Figure 1, the left half of the rectangle represents "intervention," and the right half of the rectangle represents "control." Here, the design is more complicated because the two possible randomized assignments for Monty are more complicated. The left half of the rectangle represents "reveal left if no car in left, otherwise reveal no car in right" and the right half of the rectangle represents "reveal right if no car in

right, otherwise reveal no car in left."

The data that you observe is "no car in left" and "unobserved in right." Unlike in Figure 1, you do not observe the randomized assignment (the position of the circle in the rectangle). The two rows depict the two possible joint distributions of potential outcomes that could yield the data. In the first row, in the two doors you did not choose, there is no car in left and no car in right. In the second row, in the two doors you did not choose, there is no car in left, but there is a car in right.

To decide between these two joint distributions of potential outcomes using a design-based maximum likelihood rule and the principle of maximum entropy, you calculate the likelihood in each row, depicted in the last column. In the first row, only one of the two possible randomized assignments for Monty yields the data (no car in left and unobserved in right), so the entropy is 1 and the likelihood is 1/2. In the second row, there are two randomized assignments for Monty that yield the data. (Note that in the right of the rectangle, you can observe the left potential outcome, which is not the case in the model of the experiment in Figure 1.) In the second row, the entropy is 2 and the likelihood is 1. By a design-based maximum likelihood decision rule, the maximum of the likelihood indicates that the car is behind the right remaining door, as depicted in the second row, so you should switch doors. By the principle of maximum entropy, you arrive at the same decision that you would have made under a Bayesian decision rule with a uniform prior.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association 97*(457), 284–292.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics 113*(2), 231–263.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica 88*(1), 265–296.

Alsan, M., J. Cawley, J. Doyle, Joseph J, and N. Skelley (2025, January). Mean reversion in randomized controlled trials: Implications for program targeting and heterogeneous treatment effects. Working Paper 33369, National Bureau of Economic Research.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review 112*(12), 3911–3940.

Bai, Y., S. Huang, S. Moon, A. M. Shaikh, and E. J. Vytlacil (2024). On the identifying power of monotonicity for average treatment effects. *arXiv preprint arxiv:2405.14104*.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Barnard, G. A. (1947). Significance Tests for 2 x 2 Tables. *Biometrika 34*(1-2), 123–138.

Ben-Michael, E., K. Imai, and Z. Jiang (2024). Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, 1–14.

Bernard, G. R., J.-L. Vincent, P.-F. Laterre, S. P. LaRosa, J.-F. Dhainaut, A. Lopez-Rodriguez, J. S. Steingrub, G. E. Garber, J. D. Helterbrand, E. W. Ely, and C. J. Fisher (2001). Efficacy and safety of recombinant human activated protein c for severe sepsis. *New England Journal of Medicine 344*(10), 699–709. PMID: 11236773.

Björklund, A. and R. Moffitt (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 42–49.

Boole, G. (1854). Of statistical conditions. In *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*, Chapter 19, pp. 295–319. Walton and Maberly.

Canner, P. L. (1970). Selecting one of two treatments when the responses are dichotomous. *Journal of the American Statistical Association 65*(329), 293–306.

Chan, D. C., M. Gentzkow, and C. Yu (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics 137*(2), 729–783.

Christy, N. and A. E. Kowalski (2024a). Counting defiers in health care with a design-based likelihood for the joint distribution of potential outcomes. *arXiv preprint arXiv:2412.16352d*.

Christy, N. and A. E. Kowalski (2024b). Starting small: Prioritizing safety over efficacy in randomized experiments using the exact finite sample likelihood. *arXiv preprint arxiv:2407.18206*.

Copas, J. B. (1973). Randomization models for the matched and unmatched 2 x 2 tables. *Biometrika 60*(3), 467–476.

Cox, D. R. (1958). *Planning of Experiments*. New York, NY: Wiley.

Cui, Y. and S. Han (2023). Policy learning with distributional welfare. *arXiv preprint arXiv:2311.15878*.

Dawid, A. P. and M. Musio (2022). Effects of causes and causes of effects. *Annual Review of Statistics and Its Application 9*(1), 261–287.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics 125*(1-2), 141–173.

Ding, P. and L. W. Miratrix (2019). Model-free causal inference of binary experimental data. *Scandinavian Journal of Statistics 46*(1), 200–214.

Fan, Y. and S. S. Park (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory 26*(3), 931–951.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press.

Fernández, A. A., J. L. Montiel Olea, C. Qiu, J. Stoye, and S. Tinda (2024). Robust bayes treatment choice with partial identification. *arXiv preprint arXiv:2408.11621*.

Ferrie, C. (2017, December). *Statistical physics for babies*. Baby university. Naperville, IL: Sourcebooks.

Fisher, R. (1935). *Design of Experiments* (1st ed.). Edinburgh: Oliver and Boyd.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Fréchet, M. (1957). Les tableaux de corrélation et les programmes linéaires. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute 25*(1/3), 23–40.

Freedman, D. A. and R. A. Purves (1969). Bayes' method for bookies. *The Annals of Mathematical Statistics 40*(4), 1177–1186.

Gelman, A. and G. Imbens (2013, November). Why ask why? Forward causal inference and reverse causal questions. Working Paper 19614, National Bureau of Economic Research.

Gelman, A. and K. O'Rourke (2017). Attitudes toward amalgamating evidence in statistics. https://sites.stat.columbia.edu/gelman/research/unpublished/amalgamating4.pdf.

Gneezy, U. and A. Rustichini (2000). A fine is a price. *The journal of legal studies 29*(1), 1–17.

Golan, A. (2002). Information and entropy econometrics—editor's view. *Journal of econometrics 107*(1-2), 1–15.

Greenland, S. and J. M. Robins (1986). Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology 15*(3), 413–419.

Guggenberger, P., N. Mehta, and N. Pavlov (2024). Minimax regret treatment rules with finite samples when a quantile is the object of interest. Technical report, The Pennsylvania State University.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies 64*(4), 487–535.

Heckman, J. J. and E. J. Vytlacil (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Hirano, K. (2008). Decision theory in econometrics. *The New Palgrave Dictionary of Economics, 2nd Edition. Eds. S. Durlauf and Le Blume. Palgrave Macmillan.*

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. R. Porter (2020). Asymptotic analysis of statistical decision rules in econometrics. In *Handbook of econometrics*, Volume 7, pp. 283–354. Elsevier.

Hoeffding, W. (1940). Scale-invariant correlation theory. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin 5*(3), 181–233. Translated by Dana Quade in *The Collected Works of Wassily Hoeffding*, ed. Fisher, N. I. and Sen, P. K., pp. 57–107, New York, NY: Springer New York, 1994.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Horowitz, J. L. and C. F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association 95*(449), 77–84.

Huber, M. and G. Mellace (2012, May). Relaxing monotonicity in the identification of local average treatment effects. Economics Working Paper Series 1212, University of St. Gallen, School of Economics and Political Science.

Huber, M. and G. Mellace (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics 97*(2), 398–411.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature 58*(4), 1129–1179.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Imbens, G. W. and D. B. Rubin (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies 64*(4), 555–574.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review 106*(4), 620.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. ii. *Physical review 108*(2), 171.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on systems science and cybernetics 4*(3), 227–241.

Johnson, E. J. and D. Goldstein (2003). Do defaults save lives?

Katz, L. F., J. R. Kling, J. B. Liebman, et al. (2001). Moving to opportunity in boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics 116*(2), 607–654.

Kempthorne, O. (1952). *Design and Analysis of Experiments*. New York: Wiley.

Kessler, J. B. and A. E. Roth (2025). Increasing organ donor registration as a means to increase transplantation: an experiment with actual organ donor registrations. *American Economic Journal: Economic Policy 17*(2), 60–83.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica 83*(5), 2043–2063.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kline, P. M. and C. R. Walters (2020, March). Reasonable doubt: Experimental detection of job-level employment discrimination. Working Paper 26861, National Bureau of Economic Research.

Kowalski, A. E. (2019a, March). Counting defiers. Working Paper 25671, National Bureau of Economic Research.

Kowalski, A. E. (2019b, March). A model of a randomized experiment with an application to the prowess clinical trial. Working Paper 25670, National Bureau of Economic Research.

Kowalski, A. E. (2023a). Behaviour within a clinical trial and implications for mammography guidelines. *The Review of Economic Studies 90*(1), 432–462.

Kowalski, A. E. (2023b). Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform. *The Review of Economics and Statistics 105*(3), 646–664.

Kuhn, H. W. (1953). Extensive games and the problem of information. In H. W. Kuhn and A. W. Tucker (Eds.), *Contributions to the Theory of Games, Volume II*, pp. 193–216. Princeton: Princeton University Press.

Lacetera, N. and M. Macis (2010). Do all material incentives for pro-social activities backfire? the response to cash and non-cash incentives for blood donations. *Journal of Economic Psychology 31*(4), 738–748.

Li, A. and J. Pearl (2019). Unit selection based on counterfactual logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Li, X. and P. Ding (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine 35*(6), 957–960.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics 212*, 522–555.

Manski, C. F. (1997a). The mixing problem in programme evaluation. *The Review of Economic Studies 64*(4), 537–553.

Manski, C. F. (1997b). Monotone treatment response. *Econometrica 65*(6), 1311–1334.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

Manski, C. F. (2007). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics 139*(1), 105–115.

Manski, C. F. (2018). Reasonable patient care under uncertainty. *Health Economics 27*(10), 1397–1421.

Manski, C. F. (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician 73*(sup1), 296–304.

Manski, C. F., G. D. Sandefur, S. McLanahan, and D. Powers (1992). Alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association 87*(417), 25–37.

Manski, C. F. and A. Tetenov (2007). Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference 137*(6), 1998–2010.

Manski, C. F. and A. Tetenov (2021). Statistical decision properties of imprecise trials assessing coronavirus disease 2019 (covid-19) drugs. *Value in Health 24*(5), 641–647.

Mellström, C. and M. Johannesson (2008). Crowding out in blood donation: was titmuss right? *Journal of the European Economic Association 6*(4), 845–863.

Mourifié, I. and Y. Wan (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics 99*(2), 305–313.

Mullahy, J. (2018). Individual results may vary: Inequality-probability bounds for some health-outcome treatment effects. *Journal of Health Economics 61*, 151 – 162.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Rolniczych 10*, 1–51. Translated by D.M. Dabrowski and T.P. Speed in *Statistical Science 5*(4), pp. 465–472, 1990.

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese 121*(1/2), 93–149.

Pearl, J. and D. Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. Basic books.

Permutt, T. and J. R. Hebel (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 619–622.

Richardson, T. S. and J. M. Robins (2010). Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 415–444.

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics 2*(1), 1–26.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher Randomization Test comment. *Journal of the American Statistical Association 75*(371), 591–593.

Schlag, K. H. (2003). How to minimize maximum regret in repeated decision making. *Unpublished Manuscript, European University Institute*.

Schlag, K. H. (2007). Eleven - designing randomized experiments under minimax regret. *Unpublished manuscript, European University Institute, Florence*.

Schneider, F. H., P. Campos-Mercade, S. Meier, D. Pope, E. Wengström, and A. N. Meier (2023). Financial incentives for vaccination do not have negative unintended consequences. *Nature 613*(7944), 526–533.

Semenova, V. (2024). Aggregated intersection bounds and aggregated minimax values. *arXiv preprint arXiv:2303.00982*.

Stoye, J. (2007). Minimax regret treatment choice with incomplete data and many treatments. *Econometric Theory 23*(1), 190–199.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics 151*(1), 70–81.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics 166*(1), 138–156.

Tamer, E., V. Chernozhukov, and H. Hong (2004). Parameter set inference in a class of econometric models. In *Econometric Society 2004 North American Winter Meetings*, Number 382. Econometric Society.

Tappin, D., L. Bauld, D. Purves, K. Boyd, L. Sinclair, S. MacAskill, J. McKell, B. Friel, A. McConnachie, L. De Caestecker, et al. (2015). Financial incentives for smoking cessation in pregnancy: randomised controlled trial. *Bmj 350*.

Tchetgen Tchetgen, E. J. (2024). The nudge average treatment effect. *arXiv preprint arxiv:2410.23590*.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

Tian, J. and J. Pearl (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence 28*(1-4), 287–313.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics 20*(2), 165 – 205.

Wang, J. L., T. Tran, and F. Abebe (2016). Maximum entropy and bayesian inference for the monty hall problem. *Journal of Applied Mathematics and Physics 4*(7), 1222–1230.

Welch, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika 29*(1/2), 21–52.

Young, A. (2019). Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134*(2), 557–598.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.