Testing Monotonicity in a Finite Population

Jiafeng Chen, Jonathan Roth, and Jann Spiess

ABSTRACT. We consider the extent to which we can learn from a completely randomized experiment whether all individuals have treatment effects that are weakly of the same sign, a condition we call *monotonicity*. From a classical sampling perspective, it is well-known that monotonicity is not falsifiable. By contrast, we show from the design-based perspective—in which the units in the population are fixed and only treatment assignment is stochastic—that the distribution of treatment effects in the finite population (and hence whether monotonicity holds) is formally *identified*. We argue, however, that the usual definition of identification is unnatural in the design-based setting because it imagines knowing the distribution of outcomes over different treatment assignments for the same units. We thus evaluate the informativeness of the data by the extent to which it enables frequentist testing and Bayesian updating. We show that frequentist tests can have nontrivial power against some alternatives, but power is generically limited. Likewise, we show that there exist (non-degenerate) Bayesian priors that never update about whether monotonicity holds. We conclude that, despite the formal identification result, the ability to learn about monotonicity from data in practice is severely limited.

Date: January 20, 2026. (First version: December 31, 2025.) Chen: Department of Economics, Stanford University, jiafeng@stanford.edu; Roth: Department of Economics, Brown University, jonathan_roth@brown.edu; Spiess: Graduate School of Business, Stanford University, jspiess@stanford.edu. We thank Isaiah Andrews, Neil Christy, Dean Eckles, Guido Imbens, Toru Kitagawa, Amanda Kowalski, Lihua Lei, Matt Masten, Ashesh Rambachan, Jesse Shapiro, and Kaspar Wüthrich for insightful discussions.

1. Introduction

Let D be a randomly assigned binary treatment and Y a binary outcome. Researchers are often interested in evaluating the *monotonicity* assumption that everyone has a treatment effect weakly of the same sign.[1] For example, if D is a medical intervention, we may be interested in whether some patients benefit from the treatment while others are harmed, which would indicate that outcomes could potentially be improved by better targeting. Likewise, if D is an instrumental variable and Y a treatment with imperfect compliance (more frequently denoted Z and D, respectively), then the monotonicity assumption is needed for two-stage least squares to have a local-average-treatment-effect interpretation (Angrist and Imbens, 1994).

There are two distinct approaches to statistical uncertainty in causal inference settings that lead to different formal definitions of monotonicity. The classical perspective views the n observed units as sampled from an infinite superpopulation. Monotonicity then imposes that there are no individuals in the superpopulation with opposite-signed treatment effects. Formally, we assume that the potential outcomes are sampled from a distribution g_p^*, $(y_i(1), y_i(0)) \sim g_p^*$, where $p = (p_{at}, p_{nt}, p_d, p_c)$ denotes the superpopulation shares of each "type", where each type corresponds to one of four possible values for $(y(1), y(0))$). Borrowing from the instrumental-variables literature, we use at to denote an "always-taker" with $y(0) = y(1) = 1$, and likewise use nt for never-takers ($y(0) = y(1) = 0$), d for defiers ($y(1) = 0, y(0) = 1$), and c for compliers ($y(1) = 1, y(0) = 0$). The classical definition of monotonicity is then that $\min\{p_c, p_d\} = 0$.

By contrast, the *design-based* perspective views the n units and their potential outcomes as fixed or conditioned upon, with statistical uncertainty arising only from the stochastic assignment of treatment. The finite population is characterized by $\theta = (\theta_{at}, \theta_{nt}, \theta_d, \theta_c)$, where θ_t is the *count* of how many of the n units are of type t. The design-based monotonicity assumption is then that $\min\{\theta_c, \theta_d\} = 0$.[2] In sum, the classical monotonicity assumption asks whether there are any units with opposite-signed treatment effects in the superpopulation from which the sample is drawn, while

[1] Also known as monotone treatment response (Manski, 1997).

[2] This monotonicity assumption is distinct from the one-sided hypothesis that there are no defiers (or equivalently, no negative treatment effects), which has the (generally testable) implication that the average treatment effect is non-negative. Caughey, Dafoe, Li and Miratrix (2023) employ randomization inference to construct valid test of the null of no negative treatment effects in a design-based model that also captures non-binary outcomes.

the design-based version asks whether there are any opposite-signed treatment effects *among the n units at hand*.

It is well-known that the observable data are always compatible with the null that monotonicity holds in the superpopulation. Suppose that we have a completely randomized experiment in which n_1 of the n units are randomly assigned to treatment. The researcher then observes $(Y_T, Y_U) = (\sum_i D_i Y_i, \sum_i (1 - D_i) Y_i)$, summarizing how many units in each treatment group have $Y_i = 1$. From the superpopulation perspective, we have that $(Y_T, Y_U) \sim g_p$ (where g_p is is the push-forward of g_p^* under completely random assignment with probability n_1/n). However, it is straightforward to show that p is not point-identified (see, e.g., Heckman, Smith and Clements, 1997; Fan and Park, 2010). In particular, for any p_1 that violates monotonicity, there exists a p_0 satisfying monotonicity such that $g_{p_1} = g_{p_0}$. Hence, from the superpopulation perspective, the monotonicity assumption is not falsifiable.[3]

By contrast, we show that from the design-based perspective, the type counts θ are in fact identified according to the usual technical definition of identification. Let f_θ denote the distribution of the observable data (Y_T, Y_U) over repeated assignments of treatment for the fixed finite population. The mapping $\theta \mapsto f_\theta$ is one-to-one, so θ is formally identified. At first blush, the typical definition of identification appears to suggest that the data *are* informative about whether the design-based version of monotonicity holds.

We argue, however, that the usual notion of identification is somewhat unnatural in the design-based setting: identification asks what we could learn if we saw repeated realizations of the outcomes under different treatment assignments for the *same* units. In practice, however, we only observe one realized treatment assignment for each of these units. It is therefore not clear whether identification in the usual sense implies that we can meaningfully *learn* about whether monotonicity holds from the data we actually observe. To evaluate the extent to which one can feasibly learn about monotonicity given a single realization of the data in the design-based setting, we ask two questions: First, to what extent can one construct frequentist tests of the null of monotonicity? Second, to what extent do Bayesians update about the probability that monotonicity holds?

[3] We note that monotonicity may be testable when combined with other assumptions. For example, there are tests of the joint assumptions of instrument monotonicity, exclusion, and independence (e.g. Kitagawa, 2015). In addition, monotonicity has testable implications for the marginal distributions of treated and control observations when the outcome has more than two levels (Angrist and Imbens, 1995; see also Appendix 7.2 of Caughey et al., 2023 for an application in a design-based framework).

From the frequentist perspective, we show that there exist tests of the design-based monotonicity assumption with non-trivial power against some alternatives, but we generally expect the power of these tests to be poor. In particular, we show that any test of monotonicity will have trivial or near-trivial power against some alternative: when n is even, any level-α test has power no larger than α against some alternative; when n is odd, worst-case power is bounded above by $\alpha(1 + O(2^{-n}))$. Moreover, we show that any test that has power against some alternative θ_1 will have poor power at alternatives "close" to θ_1. We formalize this by deriving upper bounds on the weighted average power (WAP) of tests in a neighborhood of θ_1, where the weights are proportional to the frequencies expected under sampling types from a superpopulation (and thus concentrate around θ_1, in terms of relative frequencies, when n is reasonably large). We show that regardless of sample size, the WAP of a size-α test is no greater than 2.51α; thus, for a 5% test, the WAP is never more than 12.6%. The bound can be even tighter given a particular n and a lower bound on the number of defiers/compliers. For example, if $n = 100$ and compliers and defiers are each at least 5% of the finite population, then WAP is never greater than 5.06%. We also derive upper bounds on power for unbiased tests (i.e. those with power at least α for all alternatives). This upper bound converges to α as n grows large, indicating that with large n all unbiased tests have nearly trivial power.

Taken together, our results suggest that frequentist tests provide limited learning about monotonicity. In particular, these tests can only be useful if we are interested in testing one particular alternative and not nearby ones. For example, with $n = 30$ and $n_1 = 15$, the highest possible power of a 5% test against any alternative is 31%, which is achieved by a test targeting the alternative with 18 defiers and 12 compliers. However, this test never rejects when there are 17 defiers and 13 compliers. Our theoretical results show the low power of this test against neighboring alternatives is a generic feature of tests against monotonicity. We suspect that in most practical settings, researchers will not be interested in ultra-specific alternatives, and thus such frequentist tests will have relatively little use in practice.

From the Bayesian perspective, we show that there do exist Bayesians who update about the probability that the design-based monotonicity assumption holds, but there also exist Bayesians who never update. The existence of a Bayesian who updates follows immediately from the fact that θ is identified. Consider a Bayesian with a two-point prior on θ_0 satisfying monotonicity and θ_1 violating monotonicity. Since $f_{\theta_0} \neq f_{\theta_1}$, the Bayesian updates about the probability that the null is true. However, we show that there exists a non-trivial prior π which never updates on the probability that the null hypothesis holds, i.e. $\pi(\theta \in \Theta_0 \mid (Y_U, Y_T)) = \pi(\theta \in \Theta_0) \in (0,1)$ (π-a.s.), where Θ_0 denotes the set of parameter values satisfying monotonicity.[4] Our results thus suggest that the data may be informative to *some* audience members, who are worried about particular violations of monotonicity, but there will not be consensus: some audience member is always unmoved by the data on monotonicity. The existence of a Bayesian who doesn't update also implies that any classifier of whether $\theta \in \Theta_0$ does no better than random guessing at some parameter value.

Related literature. Previous work has derived the likelihood of the observed data under the design-based data-generating process we consider, in which there is a completely randomized experiment with binary outcomes, and has shown that this can be used for likelihood-based or Bayesian inference (Copas, 1973; Ding and Miratrix, 2019; Christy and Kowalski, 2025). Copas (1973) and Christy and Kowalski (2025) explicitly note that the likelihood can depend on the number of defiers θ_d. We add to this literature an explicit analysis of identification, as well as results quantifying the extent to which a frequentist can test for monotonicity or a Bayesian updates about monotonicity.

More broadly, an extensive previous literature has considered the different implications of sampling-based versus model-based approaches to uncertainty for *inference* (e.g. Li and Ding, 2017; Abadie, Athey, Imbens and Wooldridge, 2020). This paper highlights that these different approaches can also have different implications for *identification*. We argue, however, that the classical notion of identification may be misleading in the design-based setting as a criterion for whether the data is informative about a parameter, and instead propose to evaluate the informativeness of the data through the properties of frequentist tests and Bayesian updating. Although we focus on testing monotonicity, these observations may prove useful in other design-based settings as well. Kline and Masten (2025) also study a design-based setting (although they do not consider monotonicity testing) and likewise find the textbook definition of identification inadequate, although they opt for defining alternative notions of identification rather than quantifying the scope for frequentist testing or Bayesian updating.

[4]Specifically, we show such priors exist when n is even. When n is odd, we show that there exist priors that *minimally* update, in the sense that the expected absolute difference between the prior and posterior probabilities for $\theta \in \Theta_0$ is $O(2^{-n})$; see Proposition A.1.

2. **Setup**

Consider a finite population of n individuals subjected to a completely randomized experiment with n_1 treated units, where $0 < n_1 < n$. That is:

(1) Each individual i has potential outcomes $(y_i(1), y_i(0)) \in \{0,1\}^2$.
(2) We observe $(Y_i, D_i)_{i=1}^n$, where $Y_i = y_i(D_i)$ and $D_i \in \{0,1\}$.
(3) Fixing the finite population, randomness solely arises from the assignment of $(D_1, \ldots, D_n)$, where

$$P(D_1 = d_1, \ldots, D_n = d_n, Y_1 = y_1, \ldots, Y_n = y_n)$$
$$= \binom{n}{n_1}^{-1} \mathbb{1}\left(\sum_{i=1}^n d_i = n_1; y_i = y_i(d_i) \text{ for all } i\right).$$

We refer to those with $y_i(1) > y_i(0)$ as compliers, $y_i(1) = 1 = y_i(0)$ as always-takers, $y_i(1) < y_i(0)$ as defiers, and $y_i(1) = 0 = y_i(0)$ as never-takers. We refer to these as the *type* of a unit. We observe the number of treated units with $Y_i = 1$ and the number of untreated units with $Y_i = 1$:[5]

$$Y := (Y_T, Y_U) = \left(\sum_{i=1}^n D_i Y_i, \sum_{i=1}^n (1 - D_i) Y_i\right).$$

The counts Y depend on the potential outcomes only through the corresponding counts of types in the finite population,

$$\theta \in \Theta := \left\{(n_{at}, n_{nt}, n_d, n_c) \in (\mathbb{N} \cup \{0\})^4 : n_{at} + n_{nt} + n_d + n_c = n\right\}.$$

In particular,

$$Y_T = (\# \text{ treated always takers}) + (\# \text{ treated compliers})$$
$$Y_U = (\# \text{ untreated always takers}) + (\# \text{ untreated defiers}).$$

In a completely randomized experiment, given $\theta = (n_{at}, n_{nt}, n_d, n_c)$, the number of treated units of each type is distributed according to a multivariate hypergeometric distribution with parameters θ and n_1. Let $f_\theta(y_T, y_U)$ denote the induced probability mass function for (Y_T, Y_U) under θ.

Sometimes, we think of the finite population as being drawn from an infinite superpopulation. To that end, let $\mathcal{P} = \{(p_{at}, p_{nt}, p_d, p_c) \in [0,1]^4 : p_{at} + p_{nt} + p_d + p_c = 1\}$

[5]We note that if the researcher observes individual data, then any procedure δ that is *anonymous*—i.e., that is invariant to permutations of the units, so that $\delta((Y_1, D_1), \ldots, (Y_n, D_n)) = \delta((Y_{\sigma(1)}, D_{\sigma(1)}), \ldots, (Y_{\sigma(n)}, D_{\sigma(n)}))$ for all permutations $\sigma : [n] \to [n]$—is simply a function of the counts (Y_T, Y_U). It thus suffices to restrict attention to (Y_T, Y_U) for anonymous procedures.

be the simplex. For an element $p \in \mathcal{P}$, let the corresponding distribution of (Y_T, Y_U) when unit types are drawn i.i.d. from Multinomial(n, p) be

$$g_p(y_T, y_U) = \sum_{\theta=(n_{at},n_{nt},n_d,n_c)\in\Theta} f_\theta(y_T, y_U)\frac{n!}{n_{at}!n_{nt}!n_c!n_d!}p_{at}^{n_{at}}p_{nt}^{n_{nt}}p_d^{n_d}p_c^{n_c}.$$

Let $\Theta_0 = \{(n_{at}, n_{nt}, n_d, n_c) \in \Theta : \min(n_d, n_c) = 0\}$ be the set of type counts that satisfy monotonicity. Let $\Theta_1 = \Theta \setminus \Theta_0$ be the complement. Similarly, let $\mathcal{P}_0 = \{p \in \mathcal{P} : \min(p_d, p_c) = 0\}$ and $\mathcal{P}_1 = \mathcal{P} \setminus \mathcal{P}_0$.

Remark 1 (Practical relevance of each null)**.** Whether we prefer to test the superpopulation null that $p \in \mathcal{P}_0$ or the design-based null that $\theta_0 \in \Theta_0$ will depend on the application. If the n units are patients in a drug trial drawn randomly from a much larger population of patients with the same condition, then we are likely more interested in whether there are heterogeneous responses to the drug in the superpopulation of patients, and thus $p \in \mathcal{P}_0$ is more relevant than $\theta \in \Theta_0$. On the other hand, if the n units are the 50 states, it may be unnatural to imagine the states as sampled from an infinite superpopulation, rendering conceptual issues for the null $p \in \mathcal{P}_0$. By contrast, testing $\theta \in \Theta_0$ answers the natural question as to whether any of the 50 states have opposite-signed treatment effects. See Copas (1973), Reichardt and Gollob (1999), and Rambachan and Roth (2025), among others, for additional discussion of the relevance of design-based vs. superpopulation-based estimands in general. For our setting specifically, Gelman and Mikhaeil (2025) argues that cases in which decision loss depends on counterfactual outcomes should be analyzed through a framework in which potential outcomes are stochastic.

3. Identification

The typical textbook definition of identification states that a parameter is (point-)identified if two distinct values of the parameter induce different distributions of the observed data.[6] It is well-known that, from the superpopulation perspective, the population shares p are not point-identified (Heckman et al., 1997). In particular, any type proportion that violates monotonicity induces data that can be rationalized by some other type proportion that obeys monotonicity. By contrast, as summarized in the following result, the type counts θ are in fact identified, and thus monotonicity violations are likewise distinguishable from f_θ.

[6]For example, the Wikipedia page on identifiability states that a statistical model P_θ is "identifiable if the mapping $\theta \mapsto P_\theta$ is one-to-one." Lehmann and Casella (1998, Definition 5.2) equivalently define θ to be unidentifiable if there exist $\theta_1 \neq \theta_2$ such that $P_{\theta_1} = P_{\theta_2}$.

Proposition 3.1. *If $\theta_1 \neq \theta_0 \in \Theta$, then $f_{\theta_1} \neq f_{\theta_0}$, and hence the finite-population type counts θ are identified. On the other hand, given any $p_1 \in \mathcal{P}_1$, there exists a $p_0 \in \mathcal{P}_0$ such that $g_{p_1} = g_{p_0}$.*

Example 1 (Illustrative example)**.** Consider a population with 2 units, one of whom is assigned to treatment. If there is 1 always-taker and 1 never-taker ($\theta = (1, 1, 0, 0)$), then (Y_T, Y_U) is either equal to $(1, 0)$ or $(0, 1)$, depending on whether the always-taker is assigned to treatment or control. By contrast, if there is 1 defier and 1 complier ($\theta = (0, 0, 1, 1)$), then (Y_T, Y_U) is either equal to $(1, 1)$ or $(0, 0)$, depending on whether the complier is assigned to treatment or control. Thus, the distribution of the observed data differs between a population with 1 always-taker and 1 never-taker and a population with 1 complier and 1 defier, despite $E[(Y_T, Y_U)]$ being the same. By contrast, a superpopulation with half always-takers and half never-takers and a superpopulation with half compliers and half defiers would each generate the same observable data distribution for (Y_T, Y_U), assigning equal probability to $(0, 0), (0, 1), (1, 0), (1, 1)$.

As a criterion for whether the observed data is informative of the parameter, however, the above definition of identification falls short from the design-based perspective. Two values of θ are distinguishable in the sense that repeated draws of (Y_T, Y_U) from f_θ would have different distributions under these two values. This, however, corresponds to knowing the distribution of outcomes from re-assigning the *same* units to *different* treatment assignments. However, in the finite population, we only observe (Y_T, Y_U) once, and thus cannot use information only learned from repeated draws. For example, knowing f_θ implies that we know $\mathrm{Cov}(Y_T, Y_U)$, yet this is difficult to learn from observing a single realization of f_θ.

In the following sections, we consider two perspectives of evaluating whether a single realization of (Y_T, Y_U) is useful for learning about monotonicity violations. From the frequentist perspective, we consider whether one can construct tests that have nontrivial power against the null of monotonicity. From the Bayesian perspective, we consider the extent to which a Bayesian updates their prior that monotonicity holds after seeing the data.

Remark 2. The discussion in previous papers sometimes suggests that θ is not identified. For example, Ding and Miratrix (2019) write that "Without monotonicity, the unknown parameters in the Science Table, $(N_{11}, N_{10}, N_{01}, N_{00})$ [θ in our notation], are no longer identifiable from the observed data." Likewise, Rosenbaum (2001) writes, "The model of a nonnegative effect cannot be verified or refuted by inspecting the responses of individuals, because r_{Ti} and r_{Ci} [$y_i(1)$ and $y_i(0)$ in our notation] are never jointly observed on the same person." Proposition 3.1 shows that according to the usual technical definition of identification, θ is in fact identified. The underlying intuition—that violations of monotonicity are hard to detect—is consistent with our power and updating results in the following sections, however.

4. **Frequentist testing**

We start by considering the possibility of frequentist testing against monotonicity. First, we show that there exist frequentist tests with power against *some* alternatives.

Proposition 4.1. (Frequentist tests exist) *Suppose $n_0, n_1 \geq 2$. Then there exist tests for monotonicity that control size and have power against some alternatives. That is, for any $\alpha \in (0,1)$ there exists a test $\delta : \mathrm{supp}(Y) \to [0,1]$ such that $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha$ and $E_{\theta_1}[\delta(Y)] > \alpha$ for some $\theta_1 \in \Theta_1$.*[7]

Intuition for tests. Observe that in Example 1, the support of Y was $\{(0,1),(1,0)\}$ when there was 1 always-taker and 1 never-taker, but was $\{(0,0),(1,1)\}$ when there was 1 complier and 1 defier. This illustrates that the support of Y may be different when monotonicity holds versus when it is violated. This general idea is the basis of the construction of tests for monotonicity in the proof to Proposition 4.1. We show that when $n \geq 4$, there always exists an alternative θ_1 such that the support of Y under θ_1 does not contain the support of Y under θ_0 for any θ_0 satisfying the null: i.e. $S^Y_{\theta_1} \not\supseteq S^Y_{\theta_0}$ for any $\theta_0 \in \Theta_0$, where S^Y_θ is the support of Y under θ.[8] It follows that a test that rejects if and only if $Y \in S^Y_{\theta_1}$ has power of 1 to reject θ_1, but only has size $\alpha_0 = \sup_{\theta_0 \in \Theta_0} E_{\theta_0}[1\{Y \in S^Y_{\theta_1}\}] < 1$. We can then construct a non-trivial randomized test with arbitrary size $\alpha \leq \alpha_0$ by rejecting with probability $\alpha/\alpha_0 > \alpha$ when $Y \in S^Y_{\theta_1}$.

Although non-trivial tests against monotonicity exist from the design-based perspective, we expect their power to be poor in practice. Our next result shows that any test of monotonicity has near-trivial power against some alternatives. Specifically, we show that if n is even, there always exists an alternative for which power is no better than size (α). For n odd, we show there exists an alternative such that power is bounded above by $\alpha(1 + O(2^{-n}))$.[9] This implies that there are no consistent tests of

[7]The proof shows that there is a non-randomized test satisfying the conditions of the proposition for α sufficiently large.

[8]Note that this is not the case in Example 1, because the support of Y when there are two always-takers is $(1,1) \subseteq S^Y_{\theta_1} = \{(0,0),(1,1)\}$. There are thus no non-trivial tests of monotonicity with $n_1 = n_0 = 1$.

[9]Surprisingly, there is a test for $n = 13, n_1 = 6$ whose minimum power over Θ_1 is ever-so-slightly larger than α. For $\alpha = 0.05$, such a test achieves power at least $0.05 + 2 \times 10^{-7}$ over all of Θ_1.

monotonicity (that is, tests for which power converges to 1 for all alternatives along a sequence of finite-populations with $n \to \infty$).

Proposition 4.2. (Near-trivial power for some alternative) *Suppose that* $\delta : \mathrm{supp}(Y) \to [0,1]$ *is a level-*α *test, i.e.* $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha$ *for some* $\alpha \in (0,1)$. *If* n *is even, then* δ *has trivial power against some alternative: there exists* $\theta_1 \in \Theta_1$ *such that* $E_{\theta_1}[\delta(Y)] \leq \alpha$. *If* n *is odd, then there exists* $\theta_1 \in \Theta_1$ *such that* $E_{\theta_1}[\delta(Y)] \leq \alpha\left(1 + \frac{1}{2^{n-1}-1}\right)$.

We can further show that the lack of power is *generic* in the following sense: if we construct a test to have power against some alternative θ_1, then there are alternatives "near" θ_1 such that power is low. More precisely, around any θ_1, we can define a weighting function that mimics sampling from a population with type frequencies $p_1 = \theta_1/n$. When n is reasonably large, these weights are "concentrated" around θ_1 in terms of type frequencies. Weighted average power (WAP) under this weighting function turns out to never be larger than 2.51α, uniformly over alternatives θ_1.

Proposition 4.3. (Low WAP) *Assume* $n \geq 4$. *Fix any* $\theta_1 \in \Theta_1$ *and let* $v = \frac{1}{n}\min(\theta_{1,c}, \theta_{1,d}) \geq 1/n$ *be the size of the monotonicity violation. Let* $\vartheta \sim \mathrm{Multinomial}(n, \theta_1/n)$. *Consider the weight function* $w_n(\cdot;\theta_1)$ *over* Θ_1 *defined by the probability mass function of* $\vartheta \mid (\vartheta \in \Theta_1)$. *Fix a test* $\delta(\cdot)$ *such that* $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha$. *Then, the weighted average power around* θ_1 *is bounded,*

$$\mathrm{WAP}(\delta;\theta_1) := \sum_{\tilde{\theta}_1 \in \Theta_1} E_{\tilde{\theta}_1}[\delta(Y)] w_n(\tilde{\theta}_1;\theta_1) \leq \alpha \frac{1}{1 - 2(1-v)^n + (1-2v)^n} \leq 2.51\alpha. \quad (1)$$

Proposition 4.3 implies, for example, that a 5% test never has weighted average power larger than 12.6%. The bound given in Proposition 4.3 becomes even tighter if one imposes a lower bound on the fraction of the finite population that are compliers/defiers. For example, in a population of 100, if $\min(\theta_c, \theta_d) \geq 5$, then the upper bound becomes 5.06%, which is virtually the same as size, uniformly over all such alternatives.

Interestingly, while we typically expect statistical power to increase with n, Proposition 4.3 implies a *tighter* upper bound on WAP the *larger* is n (holding fixed the share of compliers and defiers). Intuitively, having a larger finite population is more similar to having an infinite superpopulation, in which case there is no testable content of monotonicity.[10]

[10] Similarly, the bound on WAP gets *tighter* when the violation v of the null hypothesis is *larger*.

Proof sketch. Let $p_1 = \theta_1/n$ be the type shares under alternative θ_1. Consider a Bayesian who believes the n units to have been sampled from a superpopulation with proportions p_1, i.e. with prior $\tilde{\pi}_B \sim \text{Multinomial}(n, p_1)$. The marginal distribution over Y implied by this prior corresponds exactly to superpopulation sampling, $\tilde{\pi}_B(Y{=}y) = g_{p_1}(y)$. Now, consider a second Bayesian who believes the data to have been sampled from a superpopulation with shares p_0 satisfying monotonicity, i.e. with prior $\tilde{\pi}_A \sim \text{Multinomial}(n, p_0)$. By analogous logic, the implied marginal distribution over Y for this prior is $g_{p_0}(y)$. By Proposition 3.1, we can choose $p_0 \in \mathcal{P}_0$ such that $g_{p_0}(y) = g_{p_1}(y)$, and thus the two priors imply the same distribution over Y. It follows that the weighted average power under the two priors is the same,

$$E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)]] = E_{\theta\sim\tilde{\pi}_A}[\underbrace{E_\theta[\delta(Y)]}_{\le \alpha \text{ since } \theta\in\Theta_0}] \le \alpha,$$

where the inequality uses the fact that $\tilde{\pi}_A(\theta \in \Theta_0) = 1$ and δ controls size. Note, however, that under prior $\tilde{\pi}_B$, the design-based version of monotonicity holds with low probability: it is satisfied only if the sample of size n from a superpopulation with both compliers and defiers happens not to have zero compliers or zero defiers (or both), which occurs with probability

$$(1 - p_{1,d})^n + (1 - p_{1,c})^n - (1 - p_{1,d} - p_{1,c})^n \approx 0.$$

It follows that expected power under $\tilde{\pi}_B$ conditional on monotonicity being violated ($\theta \in \Theta_1$) is close to the unconditional expectation,

$$E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)] \mid \theta \in \Theta_1] \approx E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)]],$$

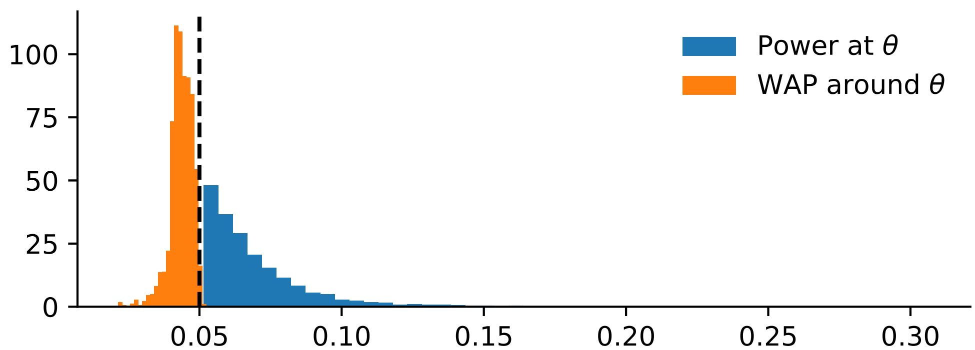
and hence

$$\text{WAP}(\delta;\theta_1) = E_{\theta\sim\pi_B}[E[\delta(Y)] \mid \theta \in \Theta_1] \approx E_{\theta\sim\pi_B}[E[\delta(Y)]] \le \alpha.$$

The proof formalizes this argument by providing an upper bound on the approximation error in the argument above, using the fact that $p_{1,d}, p_{1,c} \ge 1/n$.

Numerical illustration with $n = 30$. To illustrate these results, Figure 1 computes the most powerful 5%-level test for every possible alternative $\theta \in \Theta_1$, obtained via linear programming, for $n = 30$ and $n_1 = 15$. Perhaps surprisingly, all 4495 alternatives are testable, in the sense that for each alternative, there exists a test targeted to that alternative with power more than 0.05. However, the power for these tests for the targeted alternative tends to be modest: among these tests engineered to maximize power at a given alternative, only 53 of 4495 alternatives have tests with power above 0.15, and all of them have power below 0.31. These tests also tend to have very poor power for nearby alternatives, as suggested by Proposition 4.3: Only 18 have weighted average power above the nominal threshold 0.05, and the maximum WAP is a measly 0.0567. As a specific example, the optimal test against 18 defiers and 12 compliers achieves the maximal power of 0.31, but this test rejects with probability zero when there are 17 defiers and 13 compliers.

FIGURE 1. Power of the most powerful test for a given alternative $\theta \in \Theta_1$ for $n = 30$, $n_1 = 15$



Notes. We construct the most powerful 5%-test for a given alternative θ, which we obtain via linear programming: $\max_{\delta:\text{supp}(Y)\to[0,1]} \sum_y f_\theta(y)\delta(y)$ subjected to $\sum_y f_\vartheta(y)\delta(y) \le 0.05$ for all $\vartheta \in \Theta_0$. The blue patches show the distribution of power at θ for each test, across all 4495 values in Θ_1. For each test, we also compute its weighted average power (1) and show its distribution in orange. □

Unbiased tests. We can obtain even sharper limits on power if we restrict attention to unbiased tests. Recall that a test is unbiased if its power against all alternatives is weakly greater than its size. Proposition 4.4 below shows that unbiased tests for monotonicity do exist (at least when $n_1 = n_0$). However, Proposition 4.5 implies that unbiased tests have asymptotically trivial power.

Proposition 4.4. (Unbiased tests exist) *Suppose $n_1 = n_0 \ge 2$. Then there exists a non-trivial unbiased test of monotonicity: for any $\alpha \in (0,1)$, there exists $\delta : \text{supp}(Y) \to [0,1]$ such that $\sup_{\theta_0\in\Theta_0} E_{\theta_0}[\delta(Y)] \le \alpha \le \inf_{\theta_1\in\Theta_1} E_{\theta_1}[\delta(Y)]$, with $E_{\theta_1}[\delta(Y)] > \alpha$ for at least one $\theta_1 \in \Theta_1$.*

Proposition 4.5. (Unbiased tests have asymptotically trivial power) *Fix $\epsilon > 0$ and $n \ge 4$. Fix any $\theta \in \Theta_1$ such that $v = \min(\theta_c, \theta_d)/n \ge \epsilon$. Let δ be any unbiased level-α test, i.e. a test satisfying $\sup_{\theta_0\in\Theta_0} E_{\theta_0}[\delta(Y)] \le \alpha \le \inf_{\theta_1\in\Theta_1} E_{\theta_1}[\delta(Y)]$. Then we have* *that*

$$E_\theta[\delta(Y)] \leq \alpha(1 + \eta_n(\epsilon))$$

for $\eta_n(\epsilon) = 3.125 n^{1.5}(1-\epsilon)^n$.

In particular, Proposition 4.5 says that along any sequence of finite populations with $v_n = \min(\theta_{c,n}, \theta_{d,n})/n \geq \epsilon$, the power of any unbiased test is $(1 + o(1))\alpha$ as $n \to \infty$.

The proof of Proposition 4.5 casts the problem of maximizing $E_\theta[\delta(Y)]$ subject to unbiasedness and size control as a linear program. The dual of this program—whereby any feasible value implies an upper bound for the power of δ—involves choosing certain weights over the null and the alternative. The weights w_n defined in Proposition 4.3 turn out to enable a nontrivial upper bound of the primal value.

5. Bayesian updating

We next ask whether Bayesians update about whether $\theta \in \Theta_0$. The following result shows that *some* Bayesians update, but there exist Bayesians who find the data totally uninformative about whether monotonicity holds, in the sense that their posterior on $\theta \in \Theta_0$ is equal to their prior almost surely.

Proposition 5.1. (Bayesian updating)

(1) Some Bayesians update: there exists a prior π *with* $\pi(\theta \in \Theta_0) \in (0,1)$ *such that* $\pi(\theta \in \Theta_0 \mid Y) \neq \pi(\theta \in \Theta_0)$ *with positive* π*-probability.*

(2) There exist (nontrivial) Bayesian priors over θ *that never update about the probability that monotonicity holds: if* n *is even, then for any* $c \in (0,1)$*, there exists a prior distribution* π *over* θ *such that* $\pi(\theta \in \Theta_0 \mid Y) = \pi(\theta \in \Theta_0) = c$*,* π*-almost surely.*

We note that the conclusion in part 1 that *some* Bayesian updates is rather weak. In fact, even in the superpopulation setting, there is *some* Bayesian who updates: consider, for example, the Bayesian who believes that there are defiers if and only if the average treatment effect is larger than 0.5; this Bayesian updates about the validity of monotonicity based on the information in the data about the average treatment effect. However, the updating in the design-based setting is slightly less trivial: a Bayesian in fact updates about the relative probability of two type vectors θ_0, θ_1 that imply the same marginal distributions of $y(0)$ and $y(1)$ in the finite population, whereas this is not true in the superpopulation setting.

Part 2 of Proposition 5.1 implies that some Bayesians do not find the data informative about whether monotonicity holds at all. Hence, the data will not be persuasive to *all* audience members. This also suggests that Bayesian inference in this setting, as considered in Ding and Miratrix (2019) and Christy and Kowalski (2025), for example, will necessarily be sensitive to the choice of prior: for the class of priors in the second part of Proposition 5.1, posterior statements about monotonicity simply match the priors.

The existence of a prior that does not update also implies that any binary classifier of whether $\theta \in \Theta_0$ does no better than random guessing for some parameter value θ. Specifically, given $a \in \{0,1\}$, let $\ell(a,\theta) = 1\{a \neq 1\{\theta \in \Theta_0\}\}$ be an indicator for whether a mis-classifies whether $\theta \in \Theta_0$ (i.e. zero-one classification loss). Let $c : \mathcal{Y} \to [0,1]$ be a possibly-randomized classifier that sets $a = 1$ with probability $c(Y)$, and let $R(c,\theta) = E_\theta[c(Y)\ell(1,\theta) + (1 - c(Y)\ell(0,\theta)]$ be its average classification error at parameter θ (risk under 0-1 loss).

Corollary 5.1. *Suppose n is even. Then* $\inf_{c(\cdot)} \sup_\theta R(c,\theta) = 0.5$.

Note that a trivial classifier that guesses randomly ($c(y) = 0.5$, $\forall y$) achieves a 0.5 mis-classification rate. Corollary 5.1 implies that *every* classifier does no better than this at some parameter value θ (when n is even). This follows from Part 2 of Proposition 5.1: there exists a prior that never updates about whether $\theta \in \Theta_0$, and hence Bayes risk under this prior must be trivial. Christy and Kowalski (2025) consider the maximum likelihood estimator $\hat{\theta}$ and say that it "provides evidence in favor of monotonicity" if $\max\{\hat{\theta}_c, \hat{\theta}_d\} = 0$. Corollary 5.1 implies that this classification of whether the data support monotonicity does no better than random guessing at some parameter value.[11] Moreover, we cannot "fix" this undesirable property by choosing a different classification rule.

Part 2 of Proposition 5.1 and Corollary 5.1 focused on the case where n is even. In the appendix, we show that when n is odd, there exist non-degenerate priors that *minimally* update in the sense that the expected change between the prior and posterior is small: Proposition A.1 shows that there exists a prior such that

$$E_{Y\sim\pi_Y}[|\pi(\theta \in \Theta_0 \mid Y) - \pi(\theta \in \Theta_0)|] = O(2^{-n}),$$

[11] In fact, in some cases, this classification rule does worse than random guessing. Consider Example 1 above with $n = 2$. Note that with 1 complier and 1 defier, the support of Y is $\{(0,0),(1,1)\}$, with each support point occurring with probability 1/2. However, $Y = (1,1)$ with probability 1 if there are 2 always-takers, and $Y = (0,0)$ with probability 1 if there are two never-takers. Thus, the MLE never corresponds to having 1 complier and 1 defier, so worst-case misclassification error is 1.

where π_Y is the distribution of Y induced by prior π. This implies Corollary A.1, which states that when n is odd, a classifier of whether $\theta \in \Theta_0$ has worst-case misclassification error at least $0.5 - O(2^{-n})$.

6. Conclusion

We study what a completely randomized experiment can reveal about finite-population monotonicity. We show that from the design-based perspective, the type counts θ in the finite population are in fact identified. However, the extent to which we can feasibly learn about violations of monotonicity is severely limited: frequentist tests generically have poor power, and some Bayesians never update about whether the null is true. Thus, formal identification translates to little, if any, practical learning about monotonicity. These results highlight that conclusions about identification may differ depending on whether one adopts a sampling-based versus design-based perspective, and that studying the properties of frequentist tests and Bayesian updating may provide a more realistic assessment of the extent to which learning is possible in design-based settings. An interesting avenue for future research is to explore whether similar issues arise in other design-based causal inference problems.

Appendix A. Proofs

Proposition 3.1. *If $\theta_1 \neq \theta_0 \in \Theta$, then $f_{\theta_1} \neq f_{\theta_0}$, and hence the finite-population type counts θ are identified. On the other hand, given any $p_1 \in \mathcal{P}_1$, there exists a $p_0 \in \mathcal{P}_0$ such that $g_{p_1} = g_{p_0}$.*

Proof. For the first claim, it suffices to construct a mapping from f_θ to θ. Fix $q = n_1/n$ as the treatment probability. Observe that, by linearity of expectation,

$$E_\theta[(Y_T, Y_U)'] = (n_{at}q + n_c q, (1-q)n_{at} + (1-q)n_d)'.$$

Second, label underlying units $i = 1, \ldots, n$ and collect always takers in $N_{at} \subset [n]$, compliers in $N_c \subset [n]$, and defiers in $N_d \subset [n]$. Let $N_1 \subset [n]$ collect the (random) set of treated individuals. Observe that

$$\begin{aligned}
E_\theta[Y_T Y_U] &= \sum_{i \neq j, i \in N_{at} \cup N_c, j \in N_{at} \cup N_d} P(i \in N_1,\, j \notin N_1) && \text{(Linearity of expectations)} \\
&= |\{(i,j) \,:\, i \neq j, i \in N_{at} \cup N_c, j \in N_{at} \cup N_d\}| \cdot P(1 \in N_1,\, 2 \notin N_1) && \text{(Symmetry)} \\
&= \left[(n_{at} + n_c)(n_{at} + n_d) - n_{at}\right] P(1 \in N_1,\, 2 \notin N_1).
\end{aligned}$$

Thus, for $r = P(1 \in N_1,\, 2 \notin N_1) = \frac{n_1 n_0}{n(n-1)}$,

$$(n_{at} + n_c)(n_{at} + n_d) - n_{at} = \frac{E_\theta[Y_T Y_U]}{r)}$$

is a functional of f_θ. Letting $(\mu_T, \mu_U, \mu_{TU}) := (E_\theta[Y_T], E_\theta[Y_U], E_\theta[Y_T Y_U])$, we see that

$$\frac{\mu_T}{q} = n_{at} + n_c, \qquad \frac{\mu_U}{1-q} = n_{at} + n_d, \qquad \frac{\mu_{TU}}{r} = \frac{\mu_T}{q} \frac{\mu_U}{1-q} - n_{at}.$$

It follows that

$$n_{at} = \frac{\mu_T}{q} \frac{\mu_U}{1-q} - \frac{\mu_{TU}}{r}, \qquad n_c = \frac{\mu_T}{q} - n_{at}, \qquad n_d = \frac{\mu_U}{1-q} - n_{at}.$$

Since (μ_T, μ_U, μ_{TU}) is a function of f_θ, we have therefore shown that there is an inverse mapping from f_θ to θ, and hence the mapping $\theta \mapsto f_\theta$ must be one-to-one, as needed.

For the second claim, observe that the superpopulation data-generating process can be represented as:

(1) Fix index $i = 1, \ldots, n$. The first n_1 units are treated.
(2) For each unit i, sample her type from Multinomial$(1, p)$.
(3) The counts Y are then functions of type counts in the first n_1 units and the next $n - n_1$ units.

We can check that under this process, $\theta \sim \text{Multinomial}(n, p)$ and $Y \mid \theta \sim f_\theta$. But under this process, we have that $Y_T \sim \text{Bin}(n_1, p_{at} + p_c)$ and $Y_U \sim \text{Bin}(n - n_1, p_{at} + p_d)$, and $Y_T \perp\!\!\!\perp Y_U$. We thus see that the unconditional distribution of Y depends on p only through the sums $p_{at} + p_c$ and $p_{at} + p_d$. (Copas (1973, p. 472-3) likewise shows that the likelihood depends only on these sums.) For any given $p \in \mathcal{P}_1$, observe that p' defined by

$$p'_{at} = p_{at} + \min(p_c, p_d) \quad p'_c = p_c - \min(p_c, p_d) \quad p'_d = p_d - \min(p_c, p_d) \quad p'_{nt} = p_{nt} + \min(p_c, p_d)$$

generates observationally equivalent (Y_T, Y_U). However, $\min(p'_c, p'_d) = 0$ and thus $p' \in \mathcal{P}_0$. $\square$

Proposition 4.1. (Frequentist tests exist) *Suppose* $n_0, n_1 \geq 2$. *Then there exist tests for monotonicity that control size and have power against some alternatives. That is, for any* $\alpha \in (0,1)$ *there exists a test* $\delta : \text{supp}(Y) \to [0,1]$ *such that* $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha$ *and* $E_{\theta_1}[\delta(Y)] > \alpha$ *for some* $\theta_1 \in \Theta_1$.[12]

Proof. Suppose, without loss of generality, that $n_1 \leq n_0$ (if not, we can adopt the same argument reversing the roles of n_d and n_c). Let θ_1 be the type with $n_d = n_1 - 1$, $n_c = n - n_d$, and $n_{at} = n_{nt} = 0$. Note that $n_c = n - n_1 + 1 \geq 2$. Observe that the support of Y under θ_1 is

$$S^Y_{\theta_1} = \{(m, n_d - (n_1 - m)) : m = 1, ..., \min\{n_1, n_c\}\},$$

with $(m, n_d - (n_1 - m))$ corresponding to the realization of Y when m compliers are assigned to treatment. Observe that the points in $S^Y_{\theta_1}$ lie on an upward sloping line with slope of 1. That is, for any $y = (y_1, y_0)$ and $y' = (y'_1, y'_0)$ both in $S^Y_{\theta_1}$, we have that $y - y' = c \cdot (1, 1)$.

Now, for any $\theta_0 \in \Theta_0$, let $S^Y_{\theta_0}$ denote the support of Y under θ_0. We claim that $S^Y_{\theta_0} \not\subseteq S^Y_{\theta_1}$ for all $\theta_0 \in \Theta_0$.

Case 1 Suppose that θ_0 corresponds to there only being one type in the population. Then $S^Y_{\theta_0}$ is a singleton set with element either $(0,0)$, $(n_1, 0)$, $(0, n_0)$ (n_1, n_0) depending on whether the lone type is *nt*, *c*, *d* or *at*. It is clear that $(0,0) \notin S^Y_{\theta_1}$ and $(0, n_0) \notin S^Y_{\theta_1}$ since the first coordinate of points in $S^Y_{\theta_1}$ is strictly positive by construction. Next, note that the only possible element of $S^Y_{\theta_1}$ with first element equal to n_1 is (n_1, n_d), corresponding to the case where $m = n_1 =$

[12] The proof shows that there is a non-randomized test satisfying the conditions of the proposition for α sufficiently large.

$\min\{n_1, n_c\}$. However, by construction we have that $0 < n_d = n_1 - 1 < n_0$, and thus $(n_1, n_d) \neq (n_1, 0)$ and $(n_1, n_d) \neq (n_1, n_0)$.

Case 2 Suppose $\theta_0 \in \Theta_0$ that has positive numbers of at least two types. Suppose towards contradiction that $S^Y_{\theta_0} \subseteq S^Y_{\theta_1}$. Consider any support point (y_1, y_0) in $S^Y_{\theta_0}$. Note that $(y_1, y_0) = (\sum_i D_i y_i(1), \sum_i (1 - D_i) y_i(0))$ for some choice of D_i and a population such that $\{(y_i(1), y_i(0))\}_{i=1}^n$ has the types with frequencies given by θ_0. Since there are at least two types under θ_0, there are distinct indices j, k such that $D_j = 1, D_k = 0$ and individuals j and k are of different types, i.e. $(y_j(0), y_j(1)) \neq (y_k(0), y_k(1))$. Now consider the treatment assignment $\tilde{D}$ that swaps the treatment assignments of units j and k and leaves the other assignments unchanged, i.e. $\tilde{D}_j = 0, \tilde{D}_k = 1$, and $\tilde{D}_i = D_i$ for $i \notin \{j, k\}$. The realized outcome under treatment assignment $\tilde{D}$ is $(y_1', y_0') = (\sum_i \tilde{D}_i y_i(1), \sum_i (1 - \tilde{D}_i) y_i(0))$, and thus $(y_1', y_0') \in S^Y_{\theta_0} \subseteq S^Y_{\theta_1}$. However, we have that

$$(y_1', y_0') - (y_1, y_0) = (y_k(1) - y_j(1), y_j(0) - y_k(0)).$$

Note that since $(y_j(0), y_j(1)) \neq (y_k(0), y_k(1))$, it follows that $(y_1', y_0') - (y_1, y_0) \neq 0$. Then since $(y_1, y_0), (y_1', y_0') \in S^Y_{\theta_1}$, it must be the case that $(y_1', y_0') - (y_1, y_0) = c \cdot (1, 1)$ for an integer $c \neq 0$. It follows that

$$\begin{aligned} y_k(1) - y_j(1) &= c \\ y_j(0) - y_k(0) &= c. \end{aligned}$$

Adding the two equations, we obtain that

$$(y_k(1) - y_k(0)) - (y_j(1) - y_j(0)) = 2c \neq 0.$$

Since the treatment effects $(y_k(1) - y_k(0))$ and $(y_j(1) - y_j(0))$ are each in $\{-1, 0, 1\}$, their difference can equal an even non-zero integer only if one is equal to 1 and the other is equal to -1. However, this contradicts θ_0 satisfying monotonicity. It follows that $S^Y_{\theta_0} \not\subseteq S^Y_{\theta_1}$, as we wished to show.

Now, let $\delta^*(Y) = 1\{Y \in S^Y_{\theta_1}\}$. Since $S^Y_{\theta_0} \not\subseteq S^Y_{\theta_1}$ for all $\theta_0 \in \Theta_0$, it follows that $E_{\theta_0}[\delta^*(Y)] < 1$ for all $\theta_0 \in \Theta_0$. Since Θ_0 is finite, we thus obtain that $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta^*(Y)] < 1$. However, by construction $E_{\theta_1}[\delta^*(Y)] = 1$. Hence, $\delta^*(Y)$ controls size at level $\alpha^* = \sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta^*(Y)] < 1$ and has power of 1 against the alternative θ_1. If $\alpha^* \leq \alpha$, then the proof is complete by setting $\delta(Y) = \delta^*(Y)$ (and thus we have a non-randomized test). If $\alpha^* > \alpha$, set $\delta(Y) = \frac{\alpha}{\alpha^*}\delta^*(Y)$. Then by construction $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] = \alpha$ and $E_{\theta_1}[\delta(Y)] = \frac{\alpha}{\alpha^*} > \alpha$. $\square$

Lemma A.1. *Suppose n is even. Then there exist priors π_A and π_B on θ such that $\pi_A(\theta \in \Theta_0) = 1$, $\pi_B(\theta \in \Theta_0) = 0$, and $\pi_A(Y{=}y) = \pi_B(Y{=}y)$ for all $y \in \{0, \ldots, n_1\} \times \{0, \ldots, n_0\}$.*

Proof. Let $\tilde{\pi}_p$ denote the multinomial prior on $(n_{at}, n_{nt}, n_d, n_c)$ with probabilities $p = (p_{at}, p_{nt}, p_d, p_c)$. That is,

$$(n_{at}, n_{nt}, n_d, n_c) \sim \tilde{\pi}_p \sim \text{Multinomial}(n, p).$$

Define $\tilde{\pi}_A = \tilde{\pi}_{(0.5,0.5,0,0)}$ and $\tilde{\pi}_B = \tilde{\pi}_{(0,0,0.5,0.5)}$. The prior $\tilde{\pi}_A$ arises from assuming that the units in the finite population were sampled i.i.d. from a superpopulation in which half of individuals are always-takers and the other half are never-takers. The prior $\tilde{\pi}_B$ arises analogously if the superpopulation is half compliers and half defiers. As argued in the proof to Proposition 3.1, the marginal distribution of Y after (i) sampling $\theta \sim \text{Multinomial}(1, p)$ and (ii) generating $Y \sim f_\theta$ depends on p only through $p_{at} + p_c$ and $p_{at} + p_d$. It follows that the two priors imply the same unconditional distribution for Y, $\tilde{\pi}_A(Y{=}y) = \tilde{\pi}_B(Y{=}y)$ for all y.

Now, consider type counts $\theta_0(n_{at}) = (n_{at}, n - n_{at}, 0, 0)$, which have positive numbers of only always-takers and never-takers. Observe that the support of Y under $\theta_0(n_{at})$ is

$$S^Y_{\theta_0(n_{at})} = \{(m, n_{at} - m) \, : \, m = \max\{0, n_{at} - n_0\}, \ldots, \min\{n_1, n_{at}\}\}, \tag{2}$$

where $(m, n_{at} - m)$ is the realization of Y if m of the always-takers are selected for treatment. Hence the support points for Y under θ_0 lie on a line with slope of -1 and intercept n_{at}. Letting $S^Y_{\theta_0}$ denote the support of Y under θ_0, we see that the $S^Y_{\theta_0}$ are disjoint for all θ_0 to which $\tilde{\pi}_A$ assigns positive support:

$$S^Y_{\theta_0(n_{at})} \cap S^Y_{\theta_0(n'_{at})} = \varnothing \text{ for } n_{at} \neq n'_{at}.$$

Hence, for any set $\tilde{\Theta}_0$ of the form $\{\theta_0(n_{at,1}), \ldots, \theta_0(n_{at,K})\}$, we have that

$$\tilde{\pi}_A(Y = y \mid \theta \in \tilde{\Theta}_0) = \tilde{\pi}_A(Y = y \mid Y \in S_{\tilde{\Theta}_0}),$$

for $S_{\tilde{\Theta}_0} = \bigcup_{\theta_0 \in \tilde{\Theta}_0} S^Y_{\theta_0}$.

Similarly, let $\theta_1(n_c) = (0, 0, n - n_c, n_c)$, which has only compliers and defiers. The support points of Y under θ_1 take the form

$$S^Y_{\theta_1(n_c)} = \{(m, m + n_0 - n_c) \, : \, m = \max\{0, n_c - n_0\}, \ldots, \min\{n_1, n_c\}\}, \tag{3}$$

where $(m, n_0 - (n_c - m)) = (m, m + n_0 - n_c)$ corresponds to the realized outcomes when m compliers are selected for treatment. These points lie on an upward sloping line with slope 1 and intercept $n_0 - n_c$. Letting $S^Y_{\theta_1}$ denote the support of Y under θ_1, we again see that the $S^Y_{\theta_1}$ are disjoint. Hence, for any set $\tilde{\Theta}_1$ of the form $\{\theta_1(n_{c,1}), \ldots, \theta_1(n_{c,K})\}$, we have that

$$\tilde{\pi}_B(Y = y \mid \theta \in \tilde{\Theta}_1) = \tilde{\pi}_B(Y = y \mid Y \in S_{\tilde{\Theta}_1}),$$

for $S_{\tilde{\Theta}_1} = \bigcup_{\theta_1 \in \tilde{\Theta}_1} S^Y_{\theta_1}$.

Now, let

$$\tilde{\Theta}_0 \coloneqq \bigcup_{\substack{0 \le n_{at} \le n \\ n_{at} \equiv n_1 + 1 \pmod 2}} \{\theta_0(n_{at})\} \subset \Theta_0 \tag{4}$$

and

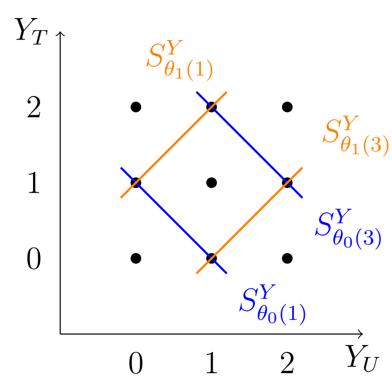
$$\tilde{\Theta}_1 \coloneqq \bigcup_{\substack{1 \le n_c \le n-1 \\ n_c \text{ odd}}} \{\theta_1(n_c)\} \subset \Theta_1. \tag{5}$$

Lemma A.2 shows formally that

$$S_{\tilde{\Theta}_1} = S_{\tilde{\Theta}_0}.$$

An intuitive illustration of the proof is shown in Figure 2 for the setting where $n_1 = n_2 = 2$, in which case $S_{\tilde{\Theta}_1}$ corresponds to the union of the points on the two upward-sloping lines in orange, and $S_{\tilde{\Theta}_0}$ corresponds the unions of the points on the two downward-sloping lines in blue.

FIGURE 2. Illustration of the sets $S_{\tilde{\Theta}_0}$ and $S_{\tilde{\Theta}_1}$ with $n_1 = n_0 = 2$



Define $\pi_A(\theta) := \tilde{\pi}_A(\theta \mid \theta \in \tilde{\Theta}_0)$ and $\pi_B(\theta) := \tilde{\pi}_B(\theta \mid \theta \in \tilde{\Theta}_1)$. It is immediate that $\pi_A(\theta \in \Theta_0) = 1$ and $\pi_B(\theta \in \Theta_1) = 1$ by construction. Further, we have shown that

$$\pi_A(Y{=}y) = \tilde{\pi}_A(Y = y \mid Y \in S_{\tilde{\Theta}_0}) = \tilde{\pi}_B(Y = y \mid Y \in S_{\tilde{\Theta}_1}) = \pi_B(Y{=}y).$$

This completes the proof.

$\square$

Lemma A.2. *Suppose n is even. Let $\tilde{\Theta}_0$, $\tilde{\Theta}_1$ be as defined in (4)-(5). Define*

$$C \coloneqq \left\{(a,b) \in \mathbb{Z}^2 :\ 0 \le a \le n_1,\ 0 \le b \le n_0,\ a+b \equiv n_1 + 1 \pmod 2\right\} \tag{6}$$

to be the set of points in $\{0,\ldots,n_1\} \times \{0,\ldots,n_0\}$ whose sum has the opposite parity of n_1. Then

$$S_{\tilde{\Theta}_0} = C = S_{\tilde{\Theta}_1}.$$

Proof. First, observe that since $n = n_0 + n_1$ is even, we have that $n_0 \equiv n_1 \pmod 2$.

Step 1: $S_{\tilde{\Theta}_0} \subseteq C$. If $(a,b) \in S_{\tilde{\Theta}_0}$ then $(a,b) \in S^Y_{\theta_0(n_{at})}$ for some n_{at} such that $n_{at} \equiv n_1 + 1 \pmod 2$. It is immediate from the expression for $S^Y_{\theta_0(n_{at})}$ in (2) that $a + b = n_{at} \equiv n_1 + 1 \pmod 2$. Moreover, by definition, $S^Y_{\theta_0(n_{at})}$ is the support of Y under $\theta_0(n_{at})$ and thus $0 \le a \le n_1$ and $0 \le b \le n_0$. Thus, $(a,b) \in C$.

Step 2: $S_{\tilde{\Theta}_1} \subseteq C$. If $(a,b) \in S_{\tilde{\Theta}_1}$, then $(a,b) \in S^Y_{\theta_1(n_c)}$ for some odd n_c such that $1 \le n_c \le n-1$. It is immediate from the expression for $S^Y_{\theta_1(n_c)}$ in (3) that $b - a = n_0 - n_c$ and hence

$$a + b \equiv b - a \equiv n_0 - n_c \pmod 2,$$

where the first $\equiv$ uses the fact that $2a \equiv 0 \pmod 2$. Using the facts that $n_0 \equiv n_1 \pmod 2$ and n_c is odd, we see that $n_0 - n_c \equiv n_0 - 1 \equiv n_1 + 1 \pmod 2$. Further, by construction $S^Y_{\theta_1(n_c)}$ is the support of Y under $\theta_1(n_c)$ and thus $0 \le a \le n_1$ and $0 \le b \le n_0$. Thus, $(a,b) \in C$.

Step 3: $C \subseteq S_{\tilde{\Theta}_0}$. Take $(a,b) \in C$ and set $n_{at} \coloneqq a + b$ and $m \coloneqq a$. Then $n_{at} \equiv n_1 + 1 \pmod 2$, and $0 \le n_{at} \le n_1 + n_0 = n$. We claim $(a,b) \in S^Y_{\theta_0(n_{at})}$. Indeed, $b \le n_0$ implies $n_{at} - n_0 = a + b - n_0 \le a$, so $m = a \ge \max(0, n_{at} - n_0)$. Also $a \le n_1$ and $a \le a + b = n_{at}$, so $m = a \le \min(n_1, n_{at})$. Thus $(a,b) = (m, n_{at} - m) \in S^Y_{\theta_0(n_{at})} \subseteq S_{\tilde{\Theta}_0}$.

Step 4: $C \subseteq S_{\tilde{\Theta}_1}$. Take $(a,b) \in C$ and set

$$m \coloneqq a, \qquad n_c \coloneqq a + n_0 - b.$$

First, $0 \le n_c \le n$ because $0 \le a \le n_1$ and $0 \le b \le n_0$ and $n_0 + n_1 = n$. Second, n_c is odd:

$$n_c = a + n_0 - b = (a - b) + n_0 \equiv (a + b) + n_0 \pmod 2,$$

where the last $\equiv$ uses the fact that $2b \equiv 0 \pmod 2$. Further, since $(a, b) \in C$, $a + b \equiv n_1 + 1 \pmod 2$, so

$$n_c \equiv (n_1 + 1) + n_0 \equiv n + 1 \equiv 1 \pmod 2,$$

where we use the fact that n is even. Finally, we check that $m = a$ lies in the admissible range for $S^Y_{\theta_1(n_c)}$. Because $n_c - n_0 = a - b$, we have $\max(0, n_c - n_0) = \max(0, a-b) \leq a = m$. Also $m = a \leq n_1$ and $m = a \leq n_c$ (since $n_c = a + n_0 - b \geq a$). Therefore $(a, b) = (m, m + n_0 - n_c) \in S^Y_{\theta_1(n_c)} \subseteq S_{\tilde{\Theta}_1}$.

Combining Steps 1–4 proves $S_{\tilde{\Theta}_0} = C = S_{\tilde{\Theta}_1}$. □

Proposition 4.2. (Near-trivial power for some alternative) *Suppose that $\delta : \mathrm{supp}(Y) \to [0, 1]$ is a level-α test, i.e. $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha$ for some $\alpha \in (0, 1)$. If n is even, then δ has trivial power against some alternative: there exists $\theta_1 \in \Theta_1$ such that $E_{\theta_1}[\delta(Y)] \leq \alpha$. If n is odd, then there exists $\theta_1 \in \Theta_1$ such that $E_{\theta_1}[\delta(Y)] \leq \alpha\left(1 + \frac{1}{2^{n-1}-1}\right)$.*

Proof. Fix some n that is even. By Lemma A.1, there exist priors π_A and π_B such that $\pi_A(\theta \in \Theta_0) = 1$, $\pi_B(\theta \in \Theta_1) = 1$, and $\pi_A(Y = y) = \pi_B(Y = y)$ for all y. Observe that

$$E_{\theta \sim \pi_A}[\delta(Y)] = E_{\theta \sim \pi_A}[\underbrace{E_\theta[\delta(Y)]}_{\leq \alpha \text{ since } \theta \in \Theta_0}] \leq \alpha,$$

where the first equality uses the law of iterated expectations, and the second the assumption that $E_\theta[\delta(Y)] \leq \alpha$ for all $\theta \in \Theta_0$. However, since $\pi_A(Y = y) = \pi_B(Y = y)$ for all y, it follows that

$$E_{\theta \sim \pi_B}[\delta(Y)] = E_{\theta \sim \pi_A}[\delta(Y)] \leq \alpha$$

and hence

$$E_{\theta \sim \pi_B}[\delta(Y)] = E_{\theta \sim \pi_B}[E_\theta[\delta(Y)]] \leq \alpha.$$

It follows that there exists some $\theta_1 \in \mathrm{supp}(\pi_B) \subseteq \Theta_1$ such that $E_\theta[\delta(Y)] \leq \alpha$, which gives the first desired result.

For the second result, fix some odd n and test δ such that $\sup_{\theta \in \Theta_0} E_\theta[\delta(Y)] \leq \alpha$. As in the proof to Lemma A.1, let $\tilde{\pi}_A$ and $\tilde{\pi}_B$ be the multinomial priors over θ with parameters $p = (0.5, 0.5, 0, 0)$ and $p = (0, 0, 0.5, 0.5)$, respectively. As argued in the proof to Lemma A.1, $\tilde{\pi}_A(Y{=}y) = \tilde{\pi}_B(Y{=}y)$ for all y. This implies that

$$E_{\theta \sim \tilde{\pi}_B}[E_\theta[\delta(Y)]] = E_{\theta \sim \tilde{\pi}_A}[E_\theta[\delta(Y)]] \leq \alpha, \tag{7}$$

where the inequality uses the fact that, by construction, $\tilde{\pi}_A(\theta \in \Theta_0) = 1$ and that δ controls size. Next, observe that $\tilde{\pi}_B(\theta \in \Theta_0) = \tilde{\pi}_B(\theta \in \{(0,0,n,0),(0,0,0,n)\} = 0.5^{n-1} =: \epsilon$. By iterated expectations, we then have

$$E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)]] = \epsilon E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)] \mid \theta \in \Theta_0] + (1-\epsilon)E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)] \mid \theta \in \Theta_1]$$

and hence

$$\begin{aligned} E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)] \mid \theta \in \Theta_1] &= \frac{1}{1-\epsilon}E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)]] - \frac{\epsilon}{1-\epsilon}E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)] \mid \theta \in \Theta_0] \\ &\le \frac{\alpha}{1-\epsilon} = \alpha\frac{1}{1-2^{-(n-1)}} = \alpha\left(1+\frac{1}{2^{n-1}-1}\right) \end{aligned}$$

where the inequality uses (7) together with the fact that $\delta(Y) \ge 0$. The result then follows from the fact that $\inf_{\theta\in\Theta_1} E_\theta[\delta(Y)] \le E_{\theta\sim\tilde{\pi}_B}[E_\theta[\delta(Y)] \mid \theta \in \Theta_1]$. ☐

Proposition 4.3. (Low WAP) *Assume $n \ge 4$. Fix any $\theta_1 \in \Theta_1$ and let $v = \frac{1}{n}\min(\theta_{1,c},\theta_{1,d}) \ge 1/n$ be the size of the monotonicity violation. Let $\vartheta \sim$ Multinomial$(n, \theta_1/n)$. Consider the weight function $w_n(\cdot;\theta_1)$ over Θ_1 defined by the probability mass function of $\vartheta \mid (\vartheta \in \Theta_1)$. Fix a test $\delta(\cdot)$ such that $\sup_{\theta_0\in\Theta_0} E_{\theta_0}[\delta(Y)] \le \alpha$. Then, the weighted average power around θ_1 is bounded,*

$$\text{WAP}(\delta;\theta_1) := \sum_{\tilde{\theta}_1\in\Theta_1} E_{\tilde{\theta}_1}[\delta(Y)]w_n(\tilde{\theta}_1;\theta_1) \le \alpha\frac{1}{1-2(1-v)^n+(1-2v)^n} \le 2.51\alpha. \quad (1)$$

Proof. Let $\pi(t;\theta_1)$ be the PMF for Multinomial$(n,\theta_1/n)$. Then

$$w_n(t;\theta_1) = \frac{\pi(t;\theta_1)}{\sum_{t\in\Theta_1}\pi(t;\theta_1)}\mathbb{1}(t\in\Theta_1).$$

Let $q = \theta_1/n$ and observe that $g_q(y) = \sum_{t\in\Theta_0\cup\Theta_1} f_t(y)\pi(t;\theta_1)$. By Proposition 3.1, there exists some $p \in \mathcal{P}_0$ such that $g_q(y) = g_p(y)$. We thus have the following expansion of WAP,

$$\begin{aligned} \text{WAP}(\delta;\theta_1) &= \sum_{t\in\Theta_0\cup\Theta_1}\sum_y \delta(y)f_t(y)\frac{\pi(t;\theta_1)}{\sum_{t\in\Theta_1}\pi(t;\theta_1)}\mathbb{1}(t\in\Theta_1) \\ &= \underbrace{\left(\sum_{t\in\Theta_1}\pi(t;\theta_1)\right)^{-1}}_{=:Z_1^{-1}}\sum_y\delta(y)\left[\sum_{t\in\Theta_0\cup\Theta_1} f_t(y)\pi(t;\theta_1) - \sum_{t\in\Theta_0} f_t(y)\pi(t;\theta_1)\right] \\ &= \frac{1}{Z_1}\left[E_{g_q}[\delta(Y)] - \sum_{t\in\Theta_0}E_t[\delta(Y)]\pi(t;\theta_1)\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{Z_1} E_{g_p}[\delta(Y)] \\
&\leq \frac{\alpha}{Z_1} && (\delta \text{ controls size})
\end{aligned}
$$

Note that

$$Z_1 = 1 - P_{V\sim\text{Multinomial}(n,q)}(\min(V_c, V_d) = 0) = 1 - \left[(1-q_c)^n + (1-q_d)^n - (1-q_c-q_d)^n\right].$$

It is easy to check that Z_1 is increasing in q_c and q_d, thus it is lower bounded:

$$Z_1 \geq 1 - 2(1-v)^n + (1-2v)^n,$$

where recall $v = \frac{1}{n}\min\{\theta_{1,c}, \theta_{1,d}\} = \min\{q_c, q_d\}$. This proves the first inequality. The second inequality follows from examining that $v \geq 1/n$ and that

$$\frac{1}{1 - 2(1-1/n)^n + (1-2/n)^n} \leq 2.51 \text{ for all } n > 3.$$

□

Proposition 4.4. (Unbiased tests exist) *Suppose $n_1 = n_0 \geq 2$. Then there exists a non-trivial unbiased test of monotonicity: for any $\alpha \in (0,1)$, there exists $\delta : \text{supp}(Y) \to [0,1]$ such that $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha \leq \inf_{\theta_1 \in \Theta_1} E_{\theta_1}[\delta(Y)]$, with $E_{\theta_1}[\delta(Y)] > \alpha$ for at least one $\theta_1 \in \Theta_1$.*

Proof. Let $\delta_0(Y) = \alpha \cdot 1\{(Y_T, Y_U) \notin \{(n_1, 0), (0, n_0)\}\}$ be the test that rejects with probability 0 if Y is either $(n_1, 0)$ or $(0, n_0)$, and rejects with probability α otherwise. Let S_θ^Y be the support of Y under parameter θ. Note that by construction, for any $\theta_0 = (n_{at}, n_{nt}, n_d, n_c) \in \Theta_0$, we have that

$$
\begin{aligned}
E_{\theta_0}[\delta_0(Y)] &< \alpha \text{ if } S_{\theta_0}^Y \cap \{(n_1, 0), (0, n_0)\} \neq \varnothing, \\
E_{\theta_0}[\delta_0(Y)] &= \alpha \text{ if } S_{\theta_0}^Y \cap \{(n_1, 0), (0, n_0)\} = \varnothing.
\end{aligned}
$$

Now, we claim that if $(n_1 - 1, 1) \in S_{\theta_0}^Y$ for $\theta_0 \in \Theta_0$, then

$$S_{\theta_0}^Y \cap \{(n_1, 0), (0, n_0)\} \neq \varnothing.$$

To see this, suppose first that θ_0 has $n_d = 0$. Then $y_i(1) \geq y_i(0)$ for all $i = 1, \ldots, n$. If there exists a treatment allocation D such that $Y(D) = (n_1 - 1, 1)$, then there exists one i for whom $(Y_i, D_i) = (1, 0)$. It follows that $y_i(0) = 1$ and hence $y_i(1) = 1$. Likewise, there must be one j for whom $(Y_j, D_j) = (0, 1)$, which implies that $Y_j(1) = 0$ and hence $Y_j(0) = 0$. Letting $\tilde{D}$ be the treatment allocation that swaps the assignments of i and j and otherwise preserves the allocation of D, we see that

$$Y(\tilde{D}) = Y(D) + (1, -1) = (n_1, 0).$$

We have thus shown that $(n_1, 0) \in S^Y_{\theta_0}$.

Similarly, suppose that θ_0 has $n_c = 0$. Then $y_i(1) \leq y_i(0)$ for $i = 1, \ldots, n$. If there exists a treatment allocation D such that $Y(D) = (n_1 - 1, 1)$, then there exists a set $A \subset [n]$ of size $n_1 - 1$ such that all $i \in A$ have $(Y_i, D_i) = (1, 1)$, which implies that $y_i(1) = 1$ and hence $y_i(0) = 1$. Likewise, there exists a set $B \subset [n]$ of size $n_0 - 1 = n_1 - 1$ such that for all $j \in B$, $(Y_j, D_j) = (0, 0)$, which implies that $Y_j(0) = 0$ and hence $Y_j(1) = 0$. Letting $\tilde{D}$ be the treatment allocation that swaps the treatment assignments of units in A and B, we see that

$$Y(\tilde{D}) = Y(D) + (-(n_1 - 1), n_1 - 1) = (0, n_1) = (0, n_0),$$

and hence $(0, n_0) \in S^Y_{\theta_0}$. This completes the proof that

$$S^Y_{\theta_0} \cap \{(n_1, 0), (0, n_0)\} \neq \varnothing.$$

Now, let $\delta_1(Y) = 1\{(Y_T, Y_U) = (n_1 - 1, 1)\}$ be the test that rejects if Y is $(n_1 - 1, 1)$. The argument above implies that for all $\theta_0 \in \Theta_0$,

$$\begin{aligned} E_{\theta_0}[\delta_0(Y)] &< \alpha \text{ if } E_{\theta_0}[\delta_1(Y)] > 0, \\ E_{\theta_0}[\delta_0(Y)] &\leq \alpha \text{ if } E_{\theta_0}[\delta_1(Y)] = 0. \end{aligned}$$

Since Θ_0 is finite, it follows that there exists $\epsilon > 0$ such that, for

$$\delta(Y) = \delta_0(Y) + \epsilon \delta_1(Y),$$

we have $E_{\theta_0}[\delta(Y)] \leq \alpha$ for all $\theta_0 \in \Theta_0$.

Next, we claim that $E_{\theta_1}[\delta(Y)] \geq \alpha$ for all $\theta_1 \in \Theta_1$. Since

$$E_{\theta_1}[\delta(Y)] \geq E_{\theta_1}[\delta_0(Y)] = \alpha\, P_{\theta_1}(Y \notin \{(n_1, 0), (0, n_0)\}),$$

it suffices to show that, for all $\theta_1 \in \Theta_1$,

$$S^Y_{\theta_1} \cap \{(n_1, 0), (0, n_0)\} = \varnothing,$$

in which case $E_{\theta_1}[\delta_0(Y)] = \alpha$. To show this, we prove the contrapositive: if S^Y_θ contains $(n_1, 0)$ or $(0, n_0)$, then $\theta \in \Theta_0$. Indeed, if there is a treatment allocation with $(Y_T, Y_U) = (n_1, 0)$, then all n_1 treated units must be always-takers or compliers, and all n_0 control units must be never-takers or compliers, and thus there can be no defiers. Analogously, if there is a treatment allocation with $(Y_T, Y_U) = (0, n_0)$, then all treated units must be never-takers or defiers, and all control units must be always-takers or defiers, and thus there are no compliers.

Finally, we show that there exists θ_1 such that $E_{\theta_1}[\delta(Y)] > \alpha$. Since we showed above that $E_{\theta_1}[\delta_0(Y)] = \alpha$ for all $\theta_1 \in \Theta_1$, it suffices to show that there exists some $\theta_1 \in \Theta_1$ such that $E_{\theta_1}[\delta_1(Y)] > 0$, or equivalently $P_{\theta_1}(Y = (n_1 - 1, 1)) > 0$. However, if θ_1 corresponds to $(n_{at}, n_{nt}, n_d, n_c) = (0, 0, 2, n-2)$, then $Y = (n_1 - 1, 1)$ obtains when one defier is assigned to treatment and the other to control. $\square$

Proposition 4.5. (Unbiased tests have asymptotically trivial power) *Fix $\epsilon > 0$ and $n \geq 4$. Fix any $\theta \in \Theta_1$ such that $v = \min(\theta_c, \theta_d)/n \geq \epsilon$. Let δ be any unbiased level-α test, i.e. a test satisfying $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[\delta(Y)] \leq \alpha \leq \inf_{\theta_1 \in \Theta_1} E_{\theta_1}[\delta(Y)]$. Then we have that*

$$E_\theta[\delta(Y)] \leq \alpha(1 + \eta_n(\epsilon))$$

for $\eta_n(\epsilon) = 3.125 n^{1.5}(1-\epsilon)^n$.

Proof. The maximum power at θ for δ is defined by the following linear program:

$$\begin{aligned} E_\theta[\delta(Y)] \leq \max_\delta \sum_y f_\theta(y)\delta(y) & \text{ subject to} \\ \sum_y f_t(y)\delta(y) &\leq \alpha \,\forall t \in \Theta_0 \\ \sum_y f_t(y)\delta(y) &\geq \alpha \,\forall t \in \Theta_1 \\ \delta(y) &\in [0,1]. \end{aligned}$$

The dual program is

$$\min_{\lambda(t)\geq 0, \mu(t) \geq 0} \alpha \left[\sum_{t\in\Theta_0} \lambda(t) - \sum_{t\in\Theta_1} \mu(t)\right] + \sum_y \left[f_\theta(y) - \sum_{t\in\Theta_0} \lambda(t) f_t(y) + \sum_{t \in \Theta_1} \mu(t) f_t(y)\right]_+ .$$

Thus, any dual feasible λ, μ implies an upper bound for $E_\theta[\delta(Y)]$.

Let $p_1 = \theta/n \in \mathcal{P}_1$ and let $\pi(t; p_1)$ be the PMF for Multinomial(n, p_1). Note that $\pi(\theta; p_1) > 0$ and define $c = \frac{1}{\pi(\theta; p_1)}$. Define, over all of $\Theta_0 \cup \Theta_1$,

$$\mu(t) = \begin{cases} c\pi(t; p_1), & t \neq \theta \\ 0 & t = \theta \end{cases}$$

Then

$$\sum_{t\in\Theta_1}\mu(t)f_t(y)+f_\theta(y)\le\sum_{t\in\Theta_1\cup\Theta_0}\mu(t)f_t(y)+f_\theta(y)=cg_{p_1}(y)=cg_{p_0}(y)$$

for some $p_0\in\mathcal{P}_0$ by Proposition 3.1. Define

$$\lambda(t)=c\pi(t;p_0)\implies\sum_{t\in\Theta_0}\lambda(t)f_t(y)=cg_{p_0}(y).$$

With this choice of λ,μ,

$$\left[f_\theta(y)-\sum_{t\in\Theta_0}\lambda(t)f_t(y)+\sum_{t\in\Theta_1}\mu(t)f_t(y)\right]_+=0.$$

Now,

$$\sum_{t\in\Theta_0}\lambda(t)-\sum_{t\in\Theta_1}\mu(t)=c-\left[c-1-c\sum_{t\in\Theta_0}\pi(t;p_1)\right]=1+\frac{\pi(\Theta_0;p_1)}{\pi(\theta;p_1)}.$$

Thus

$$E_\theta[\delta(Y)]\le\alpha\left(1+\frac{\pi(\Theta_0;p_1)}{\pi(\theta;p_1)}\right)$$

By the argument in the proof of Proposition 4.3,

$$\pi(\Theta_0;p_1)\le 2(1-v)^n-(1-2v)^n\le 2(1-\epsilon)^n.$$

Meanwhile,

$$\pi(\theta;p_1)=\frac{n!}{\theta_{at}!\theta_{nt}!\theta_c!\theta_d!}p_{1,at}^{\theta_{at}}p_{1,nt}^{\theta_{nt}}p_{1,c}^{\theta_c}p_{1,d}^{\theta_d}$$

Let $m\in\{2,3,4\}$ be the number of entries in θ that are positive. Stirling's formula implies the bounds (Robbins, 1955):

$$\sqrt{2\pi}k^{k+1/2}e^{-k}\le k!\le ek^{k+1/2}e^{-k}$$

for all integer $k\ge 1$. Plug in these bounds with $p_1=\theta/n$ to obtain

$$\pi(\theta;p_1)\ge\frac{\sqrt{2\pi}}{e^m}n^{-(m-1)/2}\prod_{j:\theta_j>0}p_{1j}^{-1/2}\ge m^{m/2}e^{-m}\sqrt{2\pi}n^{-(m-1)/2}\ge 0.64n^{-1.5}$$

where the second inequality uses the AM-GM inequality

$$\prod_{j:\theta_j>0}p_{1j}\le\left(\frac{1}{m}\sum_j p_j\right)^m=m^{-m}.$$

Therefore,

$$E_\theta[\delta(Y)] \leq \alpha(1 + 3.125n^{1.5}(1-\epsilon)^n) \to \alpha$$

as $n \to \infty$. □

Proposition 5.1. (Bayesian updating)

(1) *Some Bayesians update: there exists a prior π with $\pi(\theta \in \Theta_0) \in (0,1)$ such that $\pi(\theta \in \Theta_0 \mid Y) \neq \pi(\theta \in \Theta_0)$ with positive π-probability.*

(2) *There exist (nontrivial) Bayesian priors over θ that never update about the probability that monotonicity holds: if n is even, then for any $c \in (0,1)$, there exists a prior distribution π over θ such that $\pi(\theta \in \Theta_0 \mid Y) = \pi(\theta \in \Theta_0) = c$, π-almost surely.*

Proof. Let $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ be two parameter values. By Proposition 3.1, $f_{\theta_0} \neq f_{\theta_1}$. Let $\pi = 0.5\delta_{\theta_0} + 0.5\delta_{\theta_1}$ be such that $\pi(\Theta_0) = 0.5 \in (0,1)$. The posterior probability is such that

$$\pi(\Theta_0 \mid Y = y) = \pi(\theta_0 \mid Y = y) = 1 - \pi(\theta_1 \mid Y = y) = \frac{f_{\theta_0}(y)}{f_{\theta_1}(y) + f_{\theta_0}(y)}.$$

Since $f_{\theta_0} \neq f_{\theta_1}$ for some y in the support of one of $f_{\theta_0}(\cdot)$ and $f_{\theta_1}(\cdot)$, $\pi(\theta_0 \mid Y = y) \neq 0.5 = \pi(\theta_0)$ with positive probability. This concludes the proof of the first part.

For the second part, by Lemma A.1, there exists priors π_A and π_B such that $\pi_A(\theta \in \Theta_0) = 1$, $\pi_B(\theta \in \Theta_1)$, and $\pi_A(Y = y) = \pi_B(Y = y)$ for all y. Consider the prior $\pi_C = c\pi_A + (1-c)\pi_B$ which mixes with probabilities c and $1-c$ between π_A and π_C. By Bayes' rule, we have that

$$\pi_C(\theta \in \Theta_0 \mid Y) = \frac{\pi_C(Y \mid \theta \in \Theta_0) \cdot \pi_C(\theta \in \Theta_0)}{\pi_C(Y)}.$$

However, since $\pi_A(\theta \in \Theta_0) = 1$ and $\pi_B(\theta \in \Theta_0) = 0$, we have that $\pi_C(Y \mid \theta \in \Theta_0) = \pi_A(Y)$. Further $\pi_C(Y) = c\pi_A(Y) + (1-c)\pi_B(Y) = \pi_A(Y)$, where the second equality uses the fact that $\pi_A(Y) = \pi_B(Y)$. It follows that $\pi_C(Y \mid \theta \in \Theta_0) = \pi_A(Y) = \pi_C(Y)$. The previous display thus reduces to $\pi_C(\theta \in \Theta_0 \mid Y) = \pi_C(\theta \in \Theta_0) = c$, which gives the second result.

□

Corollary 5.1. *Suppose n is even. Then $\inf_{c(\cdot)} \sup_\theta R(c,\theta) = 0.5$.*

Proof. Observe that for any prior π over θ,

$$\inf_{c(\cdot)} \sup_\theta R(c,\theta) \geq \inf_{c(\cdot)} E_{\theta \sim \pi}[R(c,\theta)].$$

For a (possibly randomized) classifier $c(\cdot)$, it follows that:

$$\begin{aligned} E_{\theta\sim\pi}[R(c,\theta)] &= E_{\theta\sim\pi}[E_{Y\sim\pi_{Y|\theta}}[c(Y)\ell(1,\theta)+(1-c(Y)\ell(0,\theta)]] \\ &= E_{Y\sim\pi_Y}[E_{\theta\sim\pi_{\theta|Y}}[c(Y)\ell(1,\theta)+(1-c(Y)\ell(0,\theta)]] \\ &= E_{Y\sim\pi_Y}\Big[(1-\pi(\theta\in\Theta_0\mid Y))\,c(Y)+\pi(\theta\in\Theta_0\mid Y)\,(1-c(Y))\Big], \end{aligned}$$

where for clarity in the previous display, we use e.g. π_Y to denote the distribution of Y induced by the prior π, and we make explicit the measures over which random variables are taken.

By Proposition 5.1 part 2, there exists a prior π such that $\pi(\theta\in\Theta_0\mid Y)=\pi(\theta\in\Theta_0)=0.5$ (π-a.s.), in which case we have

$$E_{\theta\sim\pi}[R(c,\theta)] = E_{Y\sim\pi_Y}\Big[\tfrac{1}{2}c(Y)+\tfrac{1}{2}(1-c(Y))\Big]=\tfrac{1}{2}.$$

Hence for every classifier $c(\cdot)$, its Bayes risk under π equals $1/2$, so in particular

$$\inf_{c(\cdot)} E_{\theta\sim\pi}[R(c,\theta)]=\tfrac{1}{2},$$

and therefore

$$\inf_{c(\cdot)}\sup_{\theta} R(c,\theta)\geq\tfrac{1}{2}.$$

On the other hand, the trivial randomized classifier, $c(y)=1/2$ for all y, has

$$R(\theta,c)=\tfrac{1}{2}\quad\text{for all }\theta,$$

and thus

$$\inf_{c(\cdot)}\sup_{\theta} R(c,\theta)\leq\tfrac{1}{2}.$$

Combining the two inequalities yields $\inf_{c(\cdot)}\sup_\theta R(c,\theta)=1/2$. □

Proposition A.1. *Suppose n is odd. For any $c\in(0,1)$, there exists a prior π such that $\pi(\theta\in\Theta_0)=c$ and*

$$E_{Y\sim\pi_Y}\left[|\pi(\theta\in\Theta_0\mid Y)-\pi(\theta\in\Theta_0)|\right]=2^{-(n-2)}(1-c),$$

where π_Y is the push-forward of Y under $\theta\sim\pi$.

Proof. As in the proof to Lemma A.1, let $\tilde\pi_A$ and $\tilde\pi_B$ be the multinomial priors over θ with parameters $p=(0.5,0.5,0,0)$ and $p=(0,0,0.5,0.5)$, respectively. As argued in the proof to Lemma A.1, $\tilde\pi_A(Y{=}y)=\tilde\pi_B(Y{=}y)$ for all y. Observe that $\tilde\pi_A(\theta\in\Theta_0)=1$ and hence $\tilde\pi_A(\theta\in\Theta_0\mid Y)=1$. Next, observe that $\tilde\pi_B(\theta\in\Theta_0)=\tilde\pi_B(\theta\in\{(0,0,n,0),(0,0,0,n)\}=0.5^{n-1}=:\epsilon$. Let $C:=\{(n_1,0),(0,n_0)\}$ denote the corners of the support of Y, and observe that under $\tilde{\pi}_B$, $Y \in C$ if and only if $\theta \in \Theta_0$: that is, $\tilde{\pi}_B(Y \in C \mid \theta \in \Theta_0) = 1$, $\tilde{\pi}_B(Y \in C \mid \theta \in \Theta_1) = 0$. It follows that $\tilde{\pi}_B(Y \in C) = \epsilon$, and by Bayes' rule, $\tilde{\pi}_B(\theta \in \Theta_0 \mid Y) = 1\{Y \in C\}$.

Now, let $\pi(\theta) = \omega\tilde{\pi}_A(\theta) + (1-\omega)\tilde{\pi}_B(\theta)$ be the mixture of $\tilde{\pi}_A$ and $\tilde{\pi}_B$, with the mixture weight $\omega := \frac{c-\epsilon}{1-\epsilon}$ chosen so that $\pi(\theta \in \Theta_0) = c$. Observe that

$$\begin{aligned}
\pi(\theta \in \Theta_0 \mid Y) &= \frac{\pi(\theta \in \Theta_0, Y)}{\pi(Y)} \\
&= \frac{\omega\tilde{\pi}_A(\theta \in \Theta_0 \mid Y)\tilde{\pi}_A(Y) + (1-\omega)\tilde{\pi}_B(\theta \in \Theta_0 \mid Y)\tilde{\pi}_B(Y)}{\pi(Y)} \\
&= \omega\tilde{\pi}_A(\theta \in \Theta_0 \mid Y) + (1-\omega)\tilde{\pi}_B(\theta \in \Theta_0 \mid Y)
\end{aligned}$$

where the first line uses the definition of conditional probability, the second line uses the definition of π, and the third line uses the fact that $\tilde{\pi}_A(Y) = \tilde{\pi}_B(Y) = \pi(Y)$. It follows from the previous display and our earlier derivations of the posterior probabilities that

$$\pi(\theta \in \Theta_0 \mid Y) = \omega + (1-\omega)1\{Y \in C\}.$$

We thus see that

$$\begin{aligned}
E_{Y\sim\pi_Y}\left[|\pi(\theta \in \Theta_0 \mid Y) - \pi(\theta \in \Theta_0)|\right] &= \pi(Y \in C)(1-c) + (1-\pi(Y \in C))(c-\omega) \\
&= \epsilon\,(1-c) + (1-\epsilon)(c-\omega) \\
&= 2\epsilon(1-c).
\end{aligned}$$

The result then follows from the definition of $\epsilon = 2^{-(n-1)}$. $\square$

Corollary A.1. *Suppose n is odd. Then $\inf_{c(\cdot)}\sup_\theta R(c,\theta) \geq 0.5 - 2^{-(n-1)}$.*

Proof. As argued in the proof to Corollary 5.1, for any prior π, $\inf_{c(\cdot)}\sup_\theta R(c,\theta) \geq \inf_{c(\cdot)} E_{\theta\sim\pi}[R(c,\theta)]$ and

$$E_{\theta\sim\pi}[R(c,\theta)] = E_{Y\sim\pi_Y}\Big[(1-\pi(\theta \in \Theta_0 \mid Y))\,c(Y) + \pi(\theta \in \Theta_0 \mid Y)\,(1-c(Y)). \quad (8)$$

From Proposition A.1, there exists π such that $\pi(\theta \in \Theta_0) = 0.5$ and $E_{Y\sim\pi_Y}[|\pi(\theta \in \Theta_0 \mid Y) - \pi(\theta \in \Theta_0)|] = 2^{-(n-1)}$. Adding and subtracting $E_{Y\sim\pi_Y}[(1-\pi(\Theta_0))c(Y) + \pi(\Theta_0)(1-c(Y))] = 0.5$ from (8), we see that

$$\begin{aligned}
E_{\theta\sim\pi}[R(c,\theta)] &= 0.5 + E_{Y\sim\pi_Y}\Big[(\pi(\Theta_0 \mid Y) - \pi(\Theta_0))(1-2c(Y))\Big] \\
&\geq 0.5 - E_{Y\sim\pi_Y}[|\pi(\Theta_0 \mid Y) - \pi(\Theta_0)| \cdot |1-2c(Y)|] \\
&\geq 0.5 - E_{Y\sim\pi_Y}[|\pi(\Theta_0 \mid Y) - \pi(\Theta_0)|]
\end{aligned}$$

$$= 0.5 - 2^{-(n-1)}$$

where the second line uses the triangle inequality, and the third uses the fact that $c(Y) \in [0, 1]$. This completes the proof. $\square$

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 2020, *88* (1), 265–296. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12675.

Angrist, Joshua and Guido Imbens, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Angrist, Joshua D and Guido W Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 1995, *90* (430), 431–442.

Caughey, Devin, Allan Dafoe, Xinran Li, and Luke Miratrix, "Randomisation inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023, *85* (5), 1471–1491.

Christy, Neil and Amanda Ellen Kowalski, "Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect," December 2025. arXiv:2412.16352 [econ].

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

Ding, Peng and Luke W. Miratrix, "Model-free causal inference of binary experimental data," *Scandinavian Journal of Statistics*, 2019, *46* (1), 200–214. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12343.

Fan, Yanqin and Sang Soo Park, "SHARP BOUNDS ON THE DISTRIBUTION OF TREATMENT EFFECTS AND THEIR STATISTICAL INFERENCE," *Econometric Theory*, June 2010, *26* (3), 931–951.

Gelman, Andrew and Jonas M Mikhaeil, "Russian roulette: the need for stochastic potential outcomes when utilities depend on counterfactuals," *Biometrika*, 2025, *112* (4), asaf062.

Heckman, James J., Jeffrey Smith, and Nancy Clements, "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *The Review of Economic Studies*, October 1997, *64* (4), 487–535.

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA11974.

Kline, Brendan and Matthew A Masten, "Finite Population Identification and Design-Based Sensitivity Analysis," *arXiv preprint arXiv:2504.14127*, 2025.

Lehmann, Erich Leo and George Casella, *Theory of Point Estimation* Springer Texts in Statistics, New York: Springer-Verlag, 1998.

Li, Xinran and Peng Ding, "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, October 2017, *112* (520), 1759–1769. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1295865.

Manski, Charles F, "Monotone treatment response," *Econometrica: Journal of the Econometric Society*, 1997, pp. 1311–1334.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *Journal of the American Statistical Association*, August 2025, *0* (0), 1–15. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2025.2526700.

Reichardt, Charles S. and Harry F. Gollob, "Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample," *Psychological Methods*, 1999, *4* (1), 117–128. Place: US Publisher: American Psychological Association.

Robbins, Herbert, "A remark on Stirling's formula," *The American mathematical monthly*, 1955, *62* (1), 26–29.

Rosenbaum, Paul R., "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot," *Biometrika*, 2001, *88* (1), 219–231. Publisher: [Oxford University Press, Biometrika Trust].