Testing Mechanisms[*]

Soonwoo Kwon[†] Jonathan Roth[‡]

September 2, 2025

Abstract

Economists are often interested in the mechanisms by which a treatment affects an outcome. We develop tests for the "sharp null of full mediation" that a treatment D affects an outcome Y only through a particular mechanism (or set of mechanisms) M. Our approach exploits connections between mediation analysis and the econometric literature on testing instrument validity. We also provide tools for quantifying the magnitude of alternative mechanisms when the sharp null is rejected: we derive sharp lower bounds on the fraction of individuals whose outcome is affected by the treatment despite having the same value of M under both treatments ("always-takers"), as well as sharp bounds on the average effect of the treatment for such always-takers. An advantage of our approach relative to existing tools for mediation analysis is that it does not require stringent assumptions about how M is assigned. We illustrate our methodology in two empirical applications.

[*] We are grateful to Alberto Abadie, Don Andrews, Isaiah Andrews, Clément de Chaisemartin, Kevin Chen, Xavier D'Haultfoeuille, John Friedman, Martin Huber, Peter Hull, Toru Kitagawa, Pat Kline, Simon Lee, Caleb Miles, Ismael Mourifié, Ben Roth, Pedro Sant'Anna, Yuya Sasaki, Jesse Shapiro, Zhenting Sun, Chris Walters, Yuanyuan Wan, and participants at numerous seminars and conferences for helpful comments and suggestions. We thank Chen Cheng, Scott Lu, Jin Niu, and Eddie Wu for excellent research assistance. Roth gratefully acknowledges funding from the NIH under grant NIGMS 1R35GM155224

[†] Brown University. soonwoo_kwon@brown.edu

[‡] Brown University. jonathanroth@brown.edu

1 Introduction

Social scientists are often able to identify the causal effect of a treatment D on some outcome of interest Y, either by explicitly randomizing D or using some "quasi-experimental" variation in D. Once the causal effect of D on Y is established, a natural question is *why* does it work, i.e. what are the *mechanisms* by which D affects Y?

To fix ideas, consider the setting of Bursztyn, González and Yanagizawa-Drott (2020), which will be one of our empirical applications below. The authors show that the vast majority of men in Saudi Arabia underestimate how open other men are to women working outside of the home. They then run an experiment in which some men are randomized to receive information about other men's beliefs. At the end of the experiment, all of the men are given the choice between signing their wives up for a job-search service or taking a gift card. The authors observe that the information treatment increases the probability that men sign their wives up for the job-search service, and also increases the probability that their wives apply for and interview for jobs over the subsequent five months. A natural question in interpreting these results is then whether the increase in longer-run outcomes (e.g. job applications) is explained by the short-run sign-up for the job-search service, or whether the treatment also affects labor market outcomes through other longer-run changes in behaviors.

The literature on mediation analysis (see Huber (2019) for a review) provides formal methodology for disentangling how much of the average effect of a treatment D (e.g. information about others' beliefs) on an outcome Y (e.g. job applications) is explained by the indirect effect through some potential mediator M (e.g. job-search service sign-up). A challenge, however, is that even if the treatment D is randomly assigned, it will often be the case that the mediator of interest M is not randomly assigned.[1] Existing approaches typically make strong assumptions that allow for the identification of the causal effect of M on Y (see Related Literature below). A common assumption in the biostatistics literature, for example, is that M is as good as randomly assigned given D and some observable characteristics. This assumption will often be restrictive in applications—for example, we may worry that sign-up for the job-search service is correlated with unobservables related to women's labor supply.

In this paper, we develop methodology that sheds light on mechanisms without having to impose strong assumptions to identify the effect of M on Y. We make progress by considering an easier question than what is typically studied in the literature on mediation analysis, but one that we think will still be informative in many applications. Rather than trying to identify how much of the average effect is explained by the indirect effect through M, we start by testing what we refer to as the *sharp null of full mediation*: is the effect of D on Y fully explained through its effect

[1] One exception is "mechanism experiments" (Ludwig, Kling and Mullainathan, 2011), where the researcher explicitly randomizes an M of interest. Our focus is on the common setting where M was not randomized (e.g. due to lack of foresight, budget, or feasibility of randomization) and thus potentially endogenous.

on M? In our motivating application, the sharp null asks whether the effect of treatment on job applications is fully explained by the short-run take-up of the job-search service. More precisely, letting $Y(d,m)$ be the potential outcome as a function of treatment d and mediator m, the sharp null posits that $Y(d,m)$ depends only on m. If we can reject this null in our motivating example, then we can conclude that the treatment affects long-run outcomes through some change in behavior other than job-search service sign-up. In addition to testing the sharp null, our approach also provides useful information about the extent to which the null is violated. In particular, we develop lower bounds on the fraction of individuals whose outcome is affected by the treatment despite having the same value of M under both treatments. In our motivating example, this means we can lower bound the fraction of women whose labor market outcomes are affected by the information treatment despite the treatment having no effect on whether they sign up for the job service.

Our main theoretical results impose two assumptions. First, we suppose that the treatment D is as good as randomly assigned, i.e. D is independent of the potential outcomes $Y(\cdot,\cdot)$ and potential mediators $M(\cdot)$. In our motivating example, this is guaranteed by design since D is randomly assigned. (We consider extensions to "quasi-experimental" settings in Section 5.) Second, we allow the researcher to impose restrictions on how the mediator M responds to treatment. A leading example is the monotonicity assumption that the potential mediator $M(d)$ is increasing in d. In our motivating application, this imposes that providing men with information that other men are *more* open to women working outside of the home can only *increase* whether they sign up for the job-search service (in our main analysis, we restrict attention to the majority of men who initially underestimate others' openness, so the information plausibly updates beliefs in a common direction). We first consider the setting where monotonicity holds, and then introduce a more general framework that allows the researcher to impose arbitrary restrictions on the distribution of $(M(0),M(1))$, which nests monotonicity and relaxations thereof as special cases.

A key observation is that under the sharp null of full mediation and the independence and monotonicity assumptions just described, the treatment D is a valid instrumental variable for the local average treatment effect (LATE) of M on Y. In the case of binary D and binary M, the LATE assumptions are known to have testable implications (Balke and Pearl, 1997; Kitagawa, 2015; Huber and Mellace, 2015; Mourifié and Wan, 2017). Existing tools for testing the LATE assumptions can thus be used "off-the-shelf" for testing the sharp null of full mediation when D and M are binary, as we describe in more detail in Section 2. In our motivating example, the testable implications of the sharp null appear to be violated (significant at the 5% level), and thus we can conclude that the effect of the information treatment does not operate entirely through job-search service sign-up.

While existing tools can be used to test the sharp null in the case of a binary mediator M and a monotonicity assumption, several questions remain. First, we may be interested in testing that the treatment effect is explained by a non-binary M, or by a set of mechanisms—can the approach above be applied when M is non-binary and potentially multi-dimensional? Second, in some applications we may be concerned about violations of the monotonicity assumption—can one test the sharp null of full mediation under relaxations of this assumption? Third, if we reject the sharp null then we know that mechanisms other than M must matter, but how large is the contribution of the alternative mechanisms?

In Section 3, we develop a general framework that enables us to tackle all of these questions. We allow the mediator M to take on multiple values and to have multiple dimensions, so long as it has finite support $\{m_0,...,m_{K-1}\}$. We also allow the researcher to place arbitrary restrictions on $\theta_{lk} = P(M(0) = m_l, M(1) = m_k)$, the fraction of individuals with $M(0) = m_l$ and $M(1) = m_k$. The monotonicity assumption in the case with scalar M then corresponds to the special case where one imposes that $\theta_{lk} = 0$ if $m_l > m_k$. Our framework allows the researcher to impose weaker versions of this requirement—e.g. by allowing for up to $\bar{d}$ share of the population to be defiers—or to completely eliminate the monotonicity requirement altogether. Our framework also allows for various extensions of monotonicity to the setting with multi-dimensional M—e.g. a partial monotonicity assumption that imposes that each dimension of M is increasing in d.

We derive testable implications of the sharp null of full mediation in this general setting. These testable implications (formalized in Section 3.1) imply that for any set A and any value of the mediator m_k, the treatment effect on the compound outcome $\tilde{Y} = 1\{Y \in A, M = m_k\}$ should be no larger than the number of "compliers" with $M(0) = m_l$ and $M(1) = m_k$ for some $l \neq k$. The intuition for this is that under the sharp null, there should be no effect of the treatment on the outcome for "always-takers" with $M(1) = M(0)$. It follows that the treatment effect on $\tilde{Y}$ can only be driven by compliers, and thus the treatment effect on $\tilde{Y}$ must be weakly smaller than the number of compliers. When M is non-binary, a complication arises because the vector of shares of always-takers and compliers, denoted by θ, is only partially-identified. The testable implication is therefore that there exists *some* shares $\tilde{\theta}$ consistent with the observable data such that the inequalities described above are satisfied. Since the identified set for θ is characterized by linear inequalities, it is simple to verify whether such a $\tilde{\theta}$ exists by solving a linear program; we also show that the solution to the linear program has a closed-form solution under monotonicity. We further show that these testable implications are sharp in the sense that they exhaust all of the testable information in the data: if they are satisfied, there exists a distribution of potential outcomes (and potential mediators) consistent with the observable data such that the sharp null holds.

We also provide lower bounds on the extent to which the sharp null is violated. In particular, our results imply lower bounds on the fraction of the individuals who have $M = m_k$ under both treatments (the "k-always-takers") who are nevertheless affected by the treatment, $\nu_k = P(Y(1,m_k) \neq Y(0,m_k) \mid M(1) = M(0) = m_k)$. The lower bounds on the ν_k are informative about the prevalence of alternative mechanisms: if the lower bound on ν_k is large, then alternative mechanisms matter for a high fraction of k-always-takers. We also derive bounds on the average direct effect for k-always-takers, $ADE_k = E[Y(1,m_k) - Y(0,m_k) \mid M(1) = M(0) = m_k]$. In the special case where M is binary and one imposes monotonicity, our bounds on ADE_k match those derived in Flores and Flores-Lagunes (2010). As noted by Flores and Flores-Lagunes (2010), these bounds are equivalent to the familiar Lee (2009) bounds, treating M as the "sample selection". Our results in Section 3.2 generalize these bounds to the case where M is multi-valued and/or multi-dimensional, and allow for relaxations of monotonicity.

In Section 4, we show how one can conduct inference on the sharp null of full mediation, exploiting results from the literature on moment inequalities (Andrews, Roth and Pakes, 2023; Cox and Shi, 2022; Fang, Santos, Shaikh and Torgovitsky, 2023). In Monte Carlo simulations calibrated to our applications, we find good performance for the approach of Cox and Shi (2022), and thus recommend it in practice. Although for simplicity our main theoretical results focus on the case where D is randomized, in Section 5 we show that our results extend to other non-experimental settings, including settings with instrumental variables, conditional unconfoundedness, and distributional difference-in-differences.

We anticipate that our results will have several potential use-cases in applications, as highlighted by our empirical examples in Section 6. First, in many settings, there is an obvious mechanism by which the treatment would be expected to affect the outcome—often referred to as a "mechanical effect"—and it is of economic interest to know whether there are other mechanisms at play. Our motivating example of Bursztyn et al. (2020) is one such case, where there is the mechanical effect of the information on job applications through the job-search service, and we are interested in whether the information treatment also has an effect on other behavior outside of the lab. Our tests of the sharp null directly address whether the effect of the treatment is driven entirely by the mechanical effect: in Bursztyn et al. (2020), we reject that the impact of the information treatment on job applications is driven entirely by the mechanical effect of job-search service sign-up. This conclusion is of economic interest, since it suggests that an information treatment that was not tied to a job-search service would also have some effect on labor market outcomes. Our results also help us to quantify the magnitude of the alternative mechanisms: our lower bounds suggest that at least 11 percent of "never-takers" who would not enroll in the job-search service regardless of treatment status would nevertheless be induced to apply for jobs by receiving the information treatment (compared with an overall ATE of 0.12).

In other settings, there may not be a focal "mechanical effect", but the researcher may observe that the treatment affects a particular M (or set of M's), and conjecture that this mediator explains the treatment effect. Our tests of the sharp null, along with accompanying lower bounds on ν_k and ADE_k, help to quantify the completeness of such conjectures. This is illustrated in our second application to Baranov, Bhalotra, Biroli and Maselko (2020), who find that cognitive behavioral therapy for new mothers has an impact on women's economic outcomes. They conjecture that this effect may operate through increased presence of a grandmother in the home and improved relationship quality with the husband. Our results help us to quantify the completeness of these conjectures. Our tests reject the null hypothesis that either of these mechanisms on its own fully explains the treatment, with our lower bounds suggesting that at least 10% of always-takers are affected by the treatment for each mediator. On the other hand, we cannot statistically reject that the two mechanisms together explain the treatment effect. This, of course, does not imply that these are the only two mechanisms, but rather that the data is statistically consistent with the hypothesis that the combination of these mechanisms explains the effect.[2]

We have developed the `TestMechs` R package to facilitate implementation of the methods in this paper.

Related Literature. Our work relates to a large literature on mediation analysis. We briefly overview a few relevant strands of the literature, with a non-exhaustive list of citations, and refer the reader to VanderWeele (2016) and Huber (2019) for more comprehensive reviews. Much of the mediation analysis literature focuses on identification of average direct effects and indirect effects (e.g. Robins and Greenland, 1992; Pearl, 2001). A key challenge is that even if the treatment D is randomized, it is typically the case that the mediator M is not, and thus it is difficult to identify the effect of M on Y (conditional on D). Various strands of the literature have identified the effect of M on Y by assuming conditional unconfoundedness for M (e.g. Imai, Keele and Yamamoto, 2010), using an instrument for M (e.g. Frölich and Huber, 2017), or adopting difference-in-differences strategies (e.g. Deuchert, Huber and Schelker, 2019; Schenk, 2023). In contrast, we focus on learning about mechanisms without imposing assumptions that identify the effect of M on Y. The question we try to answer is different from most of the existing literature, however: rather than focus on average direct and indirect effects, we start by testing the *sharp null* that the effect of D on Y is fully explained by a particular mechanism (or set of mechanisms) M.[3] We further provide lower-bounds on the extent to which M does not fully explain the effect of D on Y by lower-bounding the treatment effects for always-takers who have the same value of M regardless of treatment status. We view our work as complementary to much of the literature on mediation analysis, as we impose different assumptions but also address different questions.

[2] The literature on using short-run surrogates for long-run outcomes often justifies the statistical surrogacy assumption (in part) by arguing that the treatment affects the long-run outcome only through the short-run outcome (e.g. Athey, Chetty, Imbens and Kang, 2024). In settings where both short- and long-run outcomes are available, our tests of the sharp null may also be useful for assessing the plausibility of these arguments.

[3] Miles (2023) also considers a sharp null. However, his sharp null is that either $Y(d,m)$ depends only on d *or* $M(d)$ does not depend on d, whereas we consider the sharp null that $Y(d,m)$ depends only on m. His focus is also different: rather than testing this sharp null, he considers which measures of the indirect effect are zero when his sharp null is satisfied.

A key observation in our paper is that under the sharp null of full mediation, D is an instrument for the effect of M on Y. Thus, in the setting where M is binary, existing tools for testing instrument validity with binary endogenous treatment can be used "off-the-shelf' to test the sharp null, both with monotonicity (Kitagawa, 2015; Huber and Mellace, 2015; Mourifié and Wan, 2017) and without monotonicity (Balke and Pearl, 1997; Wang, Robins and Richardson, 2017; Kédagni and Mourifié, 2020).[4] One of the key technical contributions of our paper is to derive sharp testable implications of the sharp null in the setting where M is potentially multi-valued or multi-dimensional, and where one places arbitrary restrictions on the type shares (e.g. monotonicity or relaxations thereof). Based on the equivalence between testing the sharp null and testing instrument validity described above, our results immediately imply sharp testable implications for settings with a binary instrument and multi-valued treatment, which may be of independent interest. Our testable implications build on the work of Sun (2023), who derived non-sharp testable implications of instrument validity with multi-valued treatments under monotonicity.[5]

Our paper also relates to the literature on principal stratification (Frangakis and Rubin, 2002; Zhang and Rubin, 2003; Lee, 2009). In particular, note that the sub-population of k-always takers corresponds to the so-called *principal stratum* with $M(1) = M(0) = m_k$. As noted above, in the case where M is binary, our bounds on the average effect for the always-takers match those in the aforementioned papers. Our primary focus, however, is on testing the sharp null of full mediation, which implies that the *fraction* of always-takers affected should be zero (a Fisherian sharp null), which is stronger than the weak null of a zero average effect. Moreover, the results in the literature on principal stratification typically focus on the case where M is binary, whereas our results extend to the case with multi-valued M.

Finally, we note that in empirical economics, mechanisms are often studied more informally, rather than using the formal tools for mediation discussed above. One common approach is to show the effects of D on a variety of intermediate outcomes, and to conjecture that a particular intermediate outcome M may be an important mechanism if D has an effect on M (see our application to Baranov et al., 2020 below for an example). The tools developed in this paper give formal methodology for testing the completeness of these conjectures: is the data consistent with

[4] Wang et al. (2017) consider tests of instrument validity when instrument Z, treatment D, and outcome Y are all binary, and one does not impose monotonicity. They observe that the testable implications imply lower bounds on the average controlled direct effect (ACDE) of Z on Y. Although their focus is testing instrument validity, they note in the conclusion that such lower bounds might also be used for "explaining causal mechanisms" in experiments. This observation is thus a precursor to the connections between tests for instrument validity and testing mechanisms derived in the more general setting in our paper.

[5] Another related paper is Kédagni and Mourifié (2020), who derive testable implications of instrument validity with potentially multi-valued treatments, without monotonicity. Their testable implications assume a weaker notion of independence, however, which when mapped to our context would imply that D is independent of $Y(\cdot,\cdot)$ but not $M(\cdot)$. Under this weaker notion of independence, their testable implications are sharp in the special case of binary treatment and outcome, but may not be sharp otherwise.

the hypothesized M fully explaining the treatment effect, and if not, how important are alternative mechanisms? A second common approach for evaluating mechanisms is heterogeneity analysis: is the treatment effect on Y larger in observable subgroups of the population for which the effect of D on M is larger? Although heterogeneity is often analyzed informally, this approach is sometimes formalized with an over-identification test that evaluates the null that, across subgroups defined by covariate cells, the conditional average treatment effect of D on Y is linear in the conditional average treatment effect of D on M (e.g. Angrist, Pathak and Zarate, 2023; Angrist and Hull, 2023). This approach provides a valid test of our sharp null under the additional assumption that the effect of M on Y for compliers is constant across sub-groups. By contrast, we derive testable implications of the sharp null that do not assume constant effects and do not require the presence of covariates.[6]

Set-up and Notation. Let Y denote a scalar outcome, D a binary treatment, and $M \in \mathbb{R}^p$ a p-dimensional vector of mediators with K support points, $m_0,...,m_{K-1}$. We denote by $Y(d,m)$ the potential outcome under treatment d and mediator m. Likewise, $M(d)$ denotes the potential mediator under treatment d. The researcher observes $(Y,M,D) = (Y(D,M(D)),M(D),D) \sim P_{obs}$.

2 Special Case: Binary Mediator

We first consider the special case with a binary mediator M, which helps us to develop intuition and illustrate connections to the existing literature on testing instrument validity. In the notation just introduced, this corresponds to $K = 2$, with $m_0 = 0$ and $m_1 = 1$, so that $M \in \{0,1\}$.

To fix ideas, consider the setting of Bursztyn et al. (2020). The authors conduct a randomized controlled trial (RCT) in Saudi Arabia focused on women's economic outcomes. Their analysis is motivated by the descriptive fact that at baseline in their experiment, the vast majority of men in Saudi Arabia under-estimate how open other men are to allowing women to work outside the home. After eliciting beliefs, they randomly assign a treated group of men to receive information about the other men's opinions. At the end of the experiment, both treated and untreated men choose between signing their wives up for a job-search service or taking a gift card. Bursztyn et al. (2020) find that the treatment has a positive effect on enrollment in the job-search service and on longer-run economic outcomes for women, such as applying and interviewing for jobs.

An important question in interpreting these results is whether the treatment increased long-run labor market outcomes solely by increasing take-up of the job-search service, or whether the information led men to change behavior in other ways. This question is important for understanding

[6] Moreover, our results indicate that the typical over-identification test does not exploit all the information in the data even under the assumption of constant effects: not only can one test the relationship of the average effects across covariate cells, but under the sharp null the restrictions that we derive should also hold *within* covariate cells.

what might happen if one were to provide men with information about others' beliefs without offering the opportunity to sign up for the job-search service. Bursztyn et al. (2020) write (p. 3017):

> It is difficult to separate the extent to which the longer-term effects are driven by the higher rate of access to the job service versus a persistent change in perceptions of the stigma associated with women working outside the home.

The authors provide some indirect evidence that the effects may not operate entirely through the job-search service—for example, there are effects on men's opinions in a follow-up survey—but they cannot directly link these long-run changes in opinions to economic outcomes. In what follows, we will show that in fact there is information in the data that is directly informative about the question of whether the effects on long-run labor market outcomes are driven solely by the job-search service.

For notation, let D be a binary indicator for receiving the information treatment, M a binary variable indicating job-search service sign-up, and Y a binary variable indicating applying for jobs three to five months after the experiment (i.e., a longer-term labor supply outcome). We let $Y(d,m)$ denote whether a woman would apply for jobs as a function of treatment status d and job-search service sign-up m, and let $M(d)$ denote job-search service sign-up as a function of treatment status. Since treatment is randomly assigned, it is reasonable to assume that it is independent of the potential outcomes and mediators, i.e. $D \perp\!\!\!\perp (Y(\cdot,\cdot),M(\cdot))$. For our analysis in this section, we will also impose the monotonicity assumption that receiving the information treatment weakly increases job-search service sign-up, so that $M(1) \geqslant M(0)$ (almost surely). To make this assumption reasonable, we restrict our analysis to the majority of men who prior to the experiment under-estimate other men's openness, so that all men are provided with information that other men are *more* open than they initially expected, which we expect will increase job-search service sign-up. In the subsequent sections, we will show how this monotonicity assumption can be relaxed, but imposing it will make it easier to highlight the connections to instrumental variables.

We now formalize the null hypothesis that the information treatment only affects long-run outcomes through its effect on job-search service sign-up. In particular, we say that the *sharp null of full mediation* is satisfied if

$$Y(d,m) = Y(d',m) \equiv Y(m) \text{ almost surely, for all } d,d',m \in \{0,1\}, \tag{1}$$

i.e. the treatment impacts the outcome only through its impact on M. If the sharp null holds, signing up for the job-search service is the only mechanism that matters for long-run job applications. On the other hand, if we reject the sharp null, there is evidence that other mechanisms play a role for at least some people—i.e., there is some impact of changes in beliefs on long-run outcomes that does not operate purely through sign-up for the job-search service at the end of the experiment.

Our first main observation is that if the sharp null holds (together with our assumptions of independence and monotonicity), then D is a valid instrument for the LATE of M on Y. This implies that testing the sharp null in this setting is equivalent to testing the validity of the LATE assumptions when both the treatment and instrument are binary. However, prior work has shown that in settings with a binary instrument and treatment, the LATE assumptions have testable implications (Kitagawa, 2015; Huber and Mellace, 2015; Mourifié and Wan, 2017), and thus such tools can be used to test the sharp null.[7] Applying the results in Kitagawa (2015), with M playing the role of treatment and D the role of instrument, we obtain the following sharp testable implications:

$$\begin{aligned} &P(Y \in A, M=0 \mid D=0) \geqslant P(Y \in A, M=0 \mid D=1) \text{ and} \\ &P(Y \in A, M=1 \mid D=1) \geqslant P(Y \in A, M=1 \mid D=0), \end{aligned} \tag{2}$$

for all Borel sets A.

To gain intuition for these testable restrictions, observe that under our monotonicity assumption we can divide the population into "always-takers" who enroll in the job-search service regardless of treatment $(M(0)=M(1)=1)$, "never-takers" who do not enroll regardless of treatment $(M(0)=M(1)=0)$, and "compliers" who enroll only if treated $(M(0)=0, M(1)=1)$. Now, consider the compound outcome $\tilde{Y}=1\{Y \in A, M=0\}$. For example, if $A=\{1\}$, then in our running example $\tilde{Y}$ is an indicator for the joint event of applying for a job and not signing up for the job-search service. Observe that under the sharp null, always-takers and never-takers must have the same value of $\tilde{Y}$ under both treatments: by definition, they have the same value of M under both treatments, and hence under the sharp null that D affects Y only through M, they also have the same value of Y under both treatments. Since $\tilde{Y}$ is just a compound outcome involving Y and M, it follows that they have the same value of $\tilde{Y}$ under both treatments. Observe, further, that the treatment effect of D on $\tilde{Y}$ for compliers must be weakly negative, since compliers have $M=1$ when they are treated, and thus when compliers are treated they have $\tilde{Y}=0$. It follows that under the sharp null, there must be a weakly negative treatment effect of D on $\tilde{Y}$, and hence

$$P(Y \in A, M=0 \mid D=1) - P(Y \in A, M=0 \mid D=0) \leqslant 0,$$

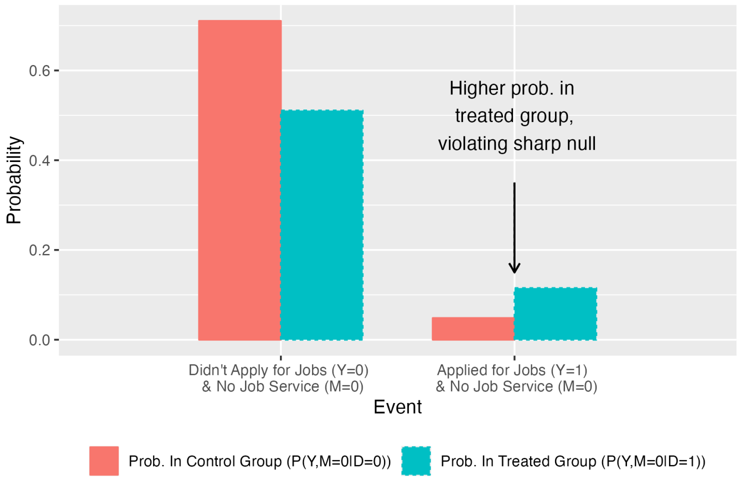
which gives the first testable implication in (2). If this implication is violated in the data, then we can conclude that—in violation of the sharp null—there must be an effect of the treatment for some never-takers.[8] The second testable implication in (2) can be derived analogously using

[7] More precisely, these tests are joint tests of the sharp null along with the independence and monotonicity assumptions. However, if we maintain that the latter two hold, then any violations must be due to violations of the sharp null. We explore relaxations of the monotonicity assumption in subsequent sections.

[8] The argument above showed that if there is a positive treatment effect on the outcome $\tilde{Y}=1\{Y=1, M=0\}$, then there must be a direct effect for either never-takers or always-takers. However, always-takers have $M=1$ under both

the compound outcome of the form $1\{Y\in A, M=1\}$.

Figure 1: Illustration of Testable Implications in Bursztyn et al. (2020)



Note: This figure shows estimates of the probabilities $P(Y=y,M=0\,|\,D=d)$ for $d=0{,}1$ and $y=0{,}1$ in the application to Bursztyn et al. (2020). For example, $P(Y=1,M=0\,|\,D=0)$ is the probability that one both applies for a job *and* does not sign up for the job-search service conditional on being in the control group. Under the sharp null of full mediation, it should be that these probabilities are higher in the control group, i.e. $P(Y=y,M=0\,|\,D=0)\geqslant P(Y=y,M=0\,|\,D=1)$ for $y=0{,}1$. We see, however, that this inequality is violated in the empirical distribution for $y=1$: more women apply for jobs and don't use the job-search service in the treated group, as indicated by the black arrow.

The argument above implies that if the sharp null holds in Bursztyn et al. (2020), there should be a negative treatment effect on the compound outcome $1\{Y=1,M=0\}$, i.e. there should be fewer women in the treated group who both apply for jobs and don't use the job service. However, as shown in Figure 1, the empirical distribution shows that the opposite is true: there are more women who apply for jobs and do not sign up for the job-search service in the treated group ($\hat{P}(Y=1,M=0\,|\,D=1)>\hat{P}(Y=1,M=0\,|\,D=0)$), indicating a violation of the sharp null. This difference is statistically significant at the 5% level, as we will describe in more detail in Section 6 after we describe methods for conducting inference.

The data thus reject the sharp null hypothesis that the impact of the information treatment on job applications operates purely through job-search service sign-up. In particular, the data suggest that some never-takers must have their outcome affected by the treatment. We can thus conclude that there is some impact of changes in beliefs on job applications that does not operate mechanically through signing up for the job-search service.

The analysis so far shows that tools originally developed for testing the LATE assumptions can be useful for testing hypotheses about mechanisms. However, several questions remain. First, our rejection of the null implies that the treatment affects the outcome through mechanisms other than

treatments, and hence always have $\tilde{Y}=0$. Thus, a positive effect on $1\{Y=1,M=0\}$ must be driven by never-takers.

job-search service sign-up, but how big are these alternative mechanisms? Second, our analysis relied on the monotonicity assumption that treatment increases job-search service sign-up, but what if we would like to relax this assumption? Third, while our motivating example had a binary M, in many cases we may be interested in testing that the treatment is explained by a non-binary mechanism, or by the combination of multiple mechanisms. Can the approach be extended to such cases?

In the subsequent section, we develop a general theoretical framework that allows us to address all of these questions. Our framework accommodates mechanisms M that are potentially multi-valued or multi-dimensional, and allows for relaxations of the monotonicity assumption. Further, in addition to deriving testable implications of the sharp null, we also derive lower bounds on the extent to which the alternative mechanisms matter—in particular, we derive bounds on the fraction of always-takers (or never-takers) that are affected by the treatment, as well as the average effect of the treatment for these always-takers.

3 Theory: General Case

We now consider the general case where M is a p-dimensional vector with finite support $\{m_0,...,m_{K-1}\}$. We denote by $G = lk$ the event that $M(0) = m_l$ and $M(1) = m_k$. We refer to individuals with $G=kk$ as the k-always takers, and individuals with $G=lk$ for $l \neq k$ as the lk-compliers. (Note that the terms "always-taker" and "complier" are used somewhat broadly here. For example, a "never-taker" in the case where M is binary would be referred to as 0-always taker, and likewise a defier would be a 10-complier.) We denote by $\theta_{lk} := P(M(0)=m_l, M(1)=m_k)$ the fraction of the population of type $G=lk$, and let $\theta \in \mathbb{R}^{K^2}$ be the vector in the (K^2-1)-dimensional simplex that collects the θ_{lk}.

Extending the definition from the previous section, we say that the sharp null of full mediation holds if

$$Y(d,m) = Y(d',m) \equiv Y(m) \text{ almost surely, for all } d,d' \in \{0,1\}, m \in \{m_0,...,m_{K-1}\}.$$

We note that if M is multi-dimensional with, say, the first dimension corresponding to mechanism A and the second corresponding to mechanism B, then the sharp null imposes that the treatment operates on Y only through its joint effect on mechanisms A and B.

For simplicity, we assume in this section that treatment assignment is independent of the potential outcomes and mediators. In Section 5, we show how the results extend to other settings, such as instrumental variables, conditional unconfoundedness, and distributional difference-in-differences.

Assumption 1 (Independence and Overlap)**.** *The treatment is independent of the potential outcomes and mediators,* $D \perp\!\!\!\perp (Y(\cdot,\cdot), M(\cdot))$*, with* $0 < P(D=1) < 1$.

For our identification results, we allow for the researcher to place arbitrary restrictions on the shares of each compliance type.

Assumption 2 (Restrictions on type shares). *$\theta \in R$ for $R \subseteq \Delta$ a closed set, where Δ denotes the (K^2-1)-dimensional simplex.*

We briefly review a few examples of restrictions on θ (i.e. choices of R) that may be natural in some applications.

Example 1 (Monotonicity and relaxations thereof).
First, consider the case where M is fully-ordered, so that $m_0 < m_1 < \ldots < m_{K-1}$. This nests the binary example from the previous section as the special case where $K=2$. Then the monotonicity assumption that $M(1) \geqslant M(0)$ corresponds to the restriction

$$R = \{\theta \in \Delta : \theta_{lk} = 0 \text{ if } l > k\}. \tag{3}$$

One could also weaken this assumption by, for example, allowing for up to $\bar{d}$ fraction of the population to be defiers, which corresponds to setting

$$R = \left\{ \theta \in \Delta : \sum_{l,k:l>k} \theta_{lk} \leqslant \bar{d} \right\}.$$

▲

Example 2 (Elementwise monotonicity).
Suppose that M is a p-dimensional vector for $p > 1$. It may sometimes be reasonable to impose that each element of $M(d)$ is increasing in d. This can be achieved by setting

$$R = \{\theta \in \Delta : \theta_{lk} = 0 \text{ if } m_l \npreceq m_k\},$$

where $m_l \preceq m_k$ if each element of m_l is less-than-or-equal to the corresponding element of m_k.[9] Similar to the previous example, one could also allow for up to $\bar{d}$ fraction of the population to have $M(0) \npreceq M(1)$. ▲

Example 3 (Bounded effect of D on M).
In some settings, it may be reasonable to impose that the treatment does not have too large an

[9] Analogous logic could be used to impose that $M(0) \preceq M(1)$ in *any* partial order, not just the elementwise one.

effect on M, at least for most people. This could be formalized by setting

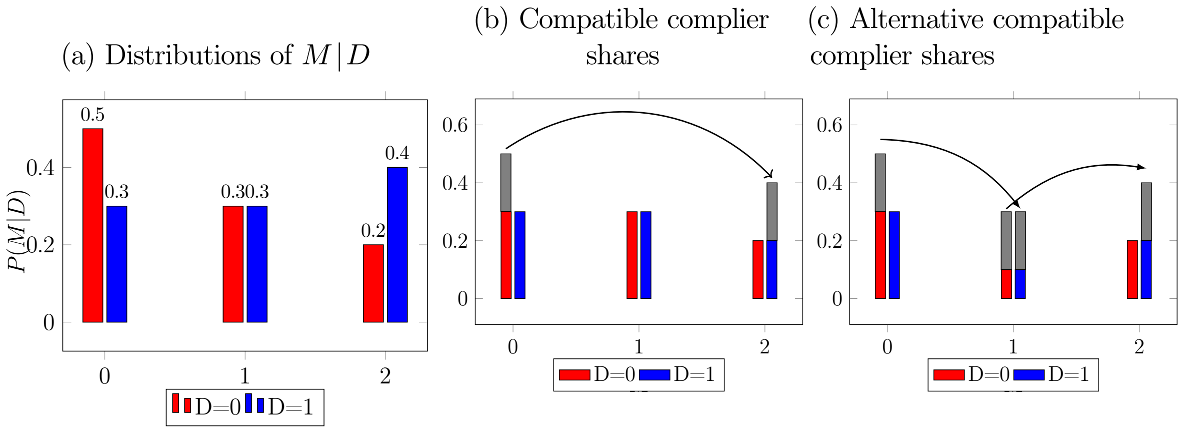
$$R = \left\{ \theta \in \Delta : \sum_{\substack{l,k \\ ||m_l - m_k|| > \kappa}} \theta_{lk} \leqslant \bar{d} \right\}.$$

This imposes that at most $\bar{d}$ fraction of the population has $||M(1) - M(0)|| > \kappa$. ▲

Example 4 (No restrictions)**.**
If the researcher is not willing to impose any restrictions on compliance types, then they can simply set $R = \Delta$. ▲

In contrast to the special case in Section 2, where the shares of each type were point-identified, in our general framework the vector of type shares θ may only be partially-identified. For example, if one relaxes the monotonicity imposed in Section 2, then analogous to the setting of instrumental variables without monotonicity (e.g. Huber, Laffers and Mellace, 2017), the share of defiers θ_{10} will generically be partially identified.[10] Partial identification of θ can also arise when M is multi-valued even if one imposes monotonicity. To see this, suppose that $M \in \{0,1,2\}$ and the marginal distributions of $M \mid D$ are as given in Figure 2, panel (a). As can be seen in the figure, the treated group has a 0.2 higher probability that $M = 2$ and a 0.2 lower probability that $M = 0$ relative to the control group. This is consistent with 20% of the population being 02-compliers and there being no other complier types (i.e. $\theta_{02} = 0.2$, $\theta_{01} = \theta_{12} = 0$), as shown in Figure 2, panel (b). However, it is also consistent with a "cascade" in which 20% of the population is 01-compliers, and another 20% of the population is 12-compliers (i.e. $\theta_{01} = \theta_{12} = 0.2$, $\theta_{02} = 0$), as shown in Figure 2, panel (c).

Figure 2: Illustration of partial identification of type shares


(a) Distributions of $M \mid D$

(b) Compatible complier shares

(c) Alternative compatible complier shares

[10] As a concrete example, suppose that $P(M = 1 \mid D = 1) = 0.5$ and $P(M = 1 \mid D = 0) = 0.3$. Then the data is consistent with there being no defiers (by setting $\theta_{11} = 0.3$, $\theta_{01} = 0.2$, $\theta_{00} = 0.5$, and $\theta_{10} = 0$) but it is also consistent with up to 0.3 fraction of the population being defiers (by setting $\theta_{11} = 0$, $\theta_{01} = 0.5$, $\theta_{00} = 0.2$, $\theta_{10} = 0.3$).

We will denote by Θ_I the set of possible values for θ (i.e. joint distributions on $(M(0),M(1))$) that are consistent with the observed distributions of $M\,|\,D$. Formally, we define the identified set Θ_I to be the set of values of $\tilde{\theta}$ such that

$$\begin{aligned}
&\sum_l \tilde{\theta}_{kl} = P(M=m_k\,|\,D=0) \text{ for } k=0,...,K-1 && \text{(Match marginals for } M\,|\,D=0\text{)}\\
&\sum_l \tilde{\theta}_{lk} = P(M=m_k\,|\,D=1) \text{ for } k=0,...,K-1 && \text{(Match marginals for } M\,|\,D=1\text{)}\\
&\tilde{\theta}\in R && \text{(Type Share Restrictions).}
\end{aligned}$$

For clarity of notation, we will use θ for the "true" shares and use $\tilde{\theta}$ to denote a generic element of the identified set Θ_I. It is worth noting that the first two restrictions above are linear in $\tilde{\theta}$. Thus, if R is characterized by linear restrictions (as is the case in Examples 1-4 above), then Θ_I is characterized by linear constraints, and thus quantities such as $\max_{\tilde{\theta}\in\Theta_I}\tilde{\theta}_{kk}$ can be calculated by linear programming. This observation will be useful for practical implementation of the testable implications below, which involve optimizations over Θ_I.

In what follows, we derive lower bounds on the extent to which the k-always takers are affected by the treatment despite having the same value of M regardless of treatment status. In particular, in Section 3.1 we derive lower bounds on the fraction of k-always takers who are affected by the treatment. These lower bounds lead naturally to tests of the sharp null of full mediation, under which the fraction of always-takers affected should be zero. In Section 3.2, we derive bounds on the average effect of the treatment for the k-always takers.

3.1 Bounds on fraction of always-takers affected

We now derive lower-bounds on the fraction of always-takers whose outcome is affected by the treatment despite having the same value of M under both treatments. To be more precise, we define

$$\nu_k := P(Y(1,m_k)\neq Y(0,m_k)\,|\,G=kk)$$

to be the fraction of k-always takers whose outcome is affected by the treatment despite always having $M=m_k$ under both treatments. The ν_k are a measure of the strength of mechanisms other than M: they tell us what fraction of the k-always takers has a direct effect of the treatment. Under the sharp null of full mediation, $Y(1,m_k)=Y(0,m_k)$ with probability 1, and thus $\nu_k=0$ for all k. By contrast, if ν_k is close to 1 for a particular k, then alternative mechanisms other than M matter for nearly all k-always takers.

Our first main result provides a lower bound on the ν_k as a function of the observable data and the type shares θ. To simplify notation, let

$$\Delta_k(A) := P(Y \in A, M = m_k \mid D=1) - P(Y \in A, M = m_k \mid D=0)$$

be the difference in the probability that $Y \in A$ and $M = m_k$ between the treated and control groups. Let $(x)_+ := \max\{x, 0\}$. We then have the following sharp lower bound on the fraction of k-always takers affected by the treatment.

Proposition 3.1.

1. ***(Lower bounds on ν_k)*** *Suppose Assumptions 1 and 2 hold. The true shares θ satisfy*

$$\theta_{kk}\nu_k \geqslant \left(\sup_A \Delta_k(A) - \sum_{l: l \neq k} \theta_{lk} \right)_+ \tag{4}$$

for $k = 0, \ldots, K-1$. Since $\theta \in \Theta_I$, it follows that there exists some $\tilde{\theta} \in \Theta_I$ such that

$$\tilde{\theta}_{kk}\nu_k \geqslant \left(\sup_A \Delta_k(A) - \sum_{l: l \neq k} \tilde{\theta}_{lk} \right)_+ \tag{5}$$

holds for all $k = 0, \ldots, K-1$.

2. ***(Sharpness)*** *The bound in (5) is sharp: for any $\tilde{\theta} \in \Theta_I$, there exists a joint distribution $P^\dagger$ for $(Y(\cdot,\cdot), M(\cdot), D)$ consistent with the observable data[11] and Assumptions 1 and 2 such that for all l and k, $P^\dagger(G = lk) = \tilde{\theta}_{lk}$ and*

$$\tilde{\theta}_{kk}\nu_k^\dagger = \left(\sup_A \Delta_k(A) - \sum_{l: l \neq k} \tilde{\theta}_{lk} \right)_+, \text{ where } \nu_k^\dagger = P^\dagger(Y(1, m_k) \neq Y(0, m_k) \mid G = kk).$$

Moreover, $P^\dagger(Y(1,m) \neq Y(0,m) \mid G = lk) = 0$ if either $l \neq k$ or $m \notin \{m_l, m_k\}$.

Equation (4) gives a lower-bound on the fraction of k-always takers whose outcome is affected by the treatment, ν_k, involving the true type shares θ and functions of the observable data (the Δ_k). The true θ will generically not be point-identified, and so (4) cannot be used directly to give a feasible lower bound on ν_k. However, we know that the true θ must lie in the identified set Θ_I. This leads to the feasible lower bound given in (5) which replaces θ in (4) with *some* $\tilde{\theta} \in \Theta_I$. The second part of Proposition 3.1 shows that the bound given in (5) is sharp in the sense that there exists a distribution of primitives consistent with the observable data such that the lower bound holds with equality. It

[11] We say that $P^\dagger$ is consistent with the observable data if $(Y(D, M(D)), M(D), D) \sim P_{obs}$ under $P^\dagger$, i.e. the distribution of realized (Y, M, D) under $P^\dagger$ matches the observed data distribution P_{obs}.

further shows that under this distribution of primitives, there is no direct effect of treatment for complier types; hence, we can only obtain non-trivial lower bounds on direct effects for the always-takers.

Recall that under the sharp null of full mediation, the fraction of always takers whose outcome is affected by the treatment should be zero. We thus immediately obtain the following testable implications of the sharp null by setting $\nu_k = 0$ in (5).

Corollary 3.1 (Testable implications of sharp null)**.**

1. ***(Testable Implications)*** *Suppose Assumptions 1 and 2 hold. If the sharp null of full mediation is satisfied, then there exists* $\tilde{\theta} \in \Theta_I$ *such that for all* $k = 0, \ldots, K-1$*,*

$$\sup_A \Delta_k(A) \leqslant \sum_{l:l \neq k} \tilde{\theta}_{lk}. \tag{6}$$

2. ***(Sharpness)*** *The testable implication in (6) is sharp: if there exists* $\tilde{\theta} \in \Theta_I$ *such that (6) holds for all* k*, then there exists a joint distribution* $P^\dagger$ *for* $(Y(\cdot,\cdot), M(\cdot), D)$ *consistent with the observable data and Assumptions 1 and 2 such that the sharp null of full mediation holds.*

Intuition. Observe that since the treatment is randomly assigned, $\Delta_k(A)$ is simply the average effect of the treatment D on the compound outcome $\tilde{Y} = 1\{Y \in A, M = m_k\}$. Note that under the sharp null of full mediation, the treatment effect of D on $\tilde{Y}$ should be zero for always-takers, since they have the same value of M and Y under both treatments. This implies that under the sharp null the effect of D on $\tilde{Y}$ is driven only by compliers and thus cannot be "too large". This is precisely what is captured in Corollary 3.1, which shows that under the sharp null, $\Delta_k(A)$ should be bounded above by the total mass of lk-compliers, $\sum_{l:l\neq k} \theta_{lk}$, regardless of the choice of A. If, in fact, the treatment effect on $\tilde{Y}$ is larger than the number of lk-compliers, then it must be that some k-always-takers had their outcome affected by the treatment, in violation of the sharp null. Indeed, (4) shows that our lower bound on the fraction of k-always-takers whose outcome is affected by the treatment, ν_k, is proportional to the positive part of the difference between $\sup_A \Delta_k(A)$ and the number of lk-compliers.

A slightly more formal sketch of the argument is as follows. We can write $\Delta_k(A) = E[\tilde{\tau}]$, where $\tilde{\tau}$ is the individual-level treatment effect of D on $\tilde{Y}$.[12] Since $\tilde{Y}$ is a binary outcome, the treatment effect $\tilde{\tau}$ must be in $\{-1, 0, 1\}$. We now argue that $\tilde{\tau}$ can equal 1 only if an individual is either an lk-complier, or a k-always taker with $Y(1, m_k) \neq Y(0, m_k)$. To see why this is a case, note that for $\tilde{\tau}$ to be 1, an individual must have $M(1) = m_k$, and thus must be either an lk-complier or a k-always taker. However, a k-always taker can have a treatment effect on $\tilde{Y}$ of 1 only if

[12] Formally, $\tilde{\tau} = \tilde{Y}(1) - \tilde{Y}(0)$ for $\tilde{Y}(d) = 1\{Y(d, M(d)) \in A, M(d) = m_k\}$.

$Y(1,m_k) \in A$ and $Y(0,m_k) \notin A$, which implies that $Y(1,m_k) \neq Y(0,m_k)$. It follows that

$$\Delta_k(A) \leqslant \underbrace{P(G=kk, Y(1,m_k) \neq Y(0,m_k))}_{\text{Prob of } k\text{-AT w/ } Y(1,m_k) \neq Y(0,m_k)} + \underbrace{\sum_{l:l\neq k} P(G=lk)}_{\text{Prob of } lk\text{-complier}}.$$

Using the fact that $\theta_{lk} = P(G=lk)$ by definition, we can rewrite the inequality as

$$\Delta_k(A) \leqslant \theta_{kk} \cdot P(Y(1,m_k) \neq Y(0,m_k) \,|\, G=kk) + \sum_{l:l\neq k} \theta_{lk}.$$

Rearranging terms, we obtain that

$$\theta_{kk} P(Y(1,m_k) \neq Y(0,m_k) \,|\, G=kk) \geqslant \Delta_k(A) - \sum_{l:l\neq k} \theta_{lk},$$

which together with the fact that probabilities are non-negative yields (4). ▲

Computation of bounds. Suppose we are interested in computing the lower-bound on ν_k for a particular k. Recall that for any $\tilde{\theta} \in \Theta_I$, we have $P(M=m_k \,|\, D=1) = \tilde{\theta}_{kk} + \sum_{l:l\neq k} \tilde{\theta}_{lk}$. It follows that we can re-write (5) as

$$\tilde{\theta}_{kk} \nu_k \geqslant \left(\sup_A \Delta_k(A) - (P(M=m_k \,|\, D=1) - \tilde{\theta}_{kk}) \right)_+,$$

where now the lower-bound depends on $\tilde{\theta}$ only through $\tilde{\theta}_{kk}$. To compute a lower-bound on ν_k, we must minimize the lower-bound for ν_k given in the previous display over $\tilde{\theta} \in \Theta_I$. It can be shown (see Lemma B.1) that the minimum is actually obtained at the minimum possible value of $\tilde{\theta}_{kk}$, i.e. by plugging in $\tilde{\theta}_{kk}^{min} := \inf_{\tilde{\theta} \in \Theta_I} \tilde{\theta}_{kk}$ into the expression in the previous display. When R is a polyhedron (as in our examples above), the identified set Θ_I is characterized by linear inequalities, and thus $\tilde{\theta}_{kk}^{min}$ can be easily computed by solving a linear program. Assuming $\tilde{\theta}_{kk}^{min} > 0$, we then obtain the bound

$$\nu_k \geqslant \frac{1}{\tilde{\theta}_{kk}^{min}} \left(\sup_A \Delta_k(A) - (P(M=m_k \,|\, D=1) - \tilde{\theta}_{kk}^{min}) \right)_+.$$

Similarly, to test whether the observable data is compatible with the sharp null we must verify whether there is any $\tilde{\theta} \in \Theta_I$ such that $\sup_A \Delta_k(A) \leqslant \sum_{l:l\neq k} \tilde{\theta}_{lk}$ for all k. By the same argument as in the previous paragraph, this is equivalent to testing whether there is any $\tilde{\theta} \in \Theta_I$ such that $\sup_A \Delta_k(A) \leqslant P(M=m_k \,|\, D=1) - \tilde{\theta}_{kk}$ for all k. Such a $\tilde{\theta} \in \Theta_I$ exists if and only if the solution to the linear program

$$\min_{s\in\mathbb{R},\tilde{\theta}\in\Theta_I} s \text{ s.t. } \sup_A \Delta_k(A) \leqslant P(M=m_k \,|\, D=1) - \tilde{\theta}_{kk} + s \text{ for all } k \tag{7}$$

is weakly negative, and so given knowledge of the distribution of the observable data, testing the implications of the sharp null is equivalent to solving a linear program.[13]

Remark 1 (Closed-form solution with fully-ordered, monotone M)**.**
Consider the case where M is fully-ordered and we impose monotonicity as in Example 1. In this case, it turns out that there is a closed-form solution for $\tilde{\theta}^{min}_{kk}$. Intuitively, to minimize the number of always-takers, we wish to have as many compliers as possible. This can be achieved by maximizing the amount of "cascading", as in panel (c) of Figure 2. Lemma B.2 in the appendix formalizes this intuition, and shows that

$$\tilde{\theta}^{min}_{kk} = \max\{\underbrace{P(M=m_k \,|\, D=1)}_{\text{Point mass at } M=m_k \text{ when } D=1} - \underbrace{(P(M\geqslant m_k \,|\, D=1) - P(M\geqslant m_k \,|\, D=0))}_{\text{Treatment effect on survival fn of } M \text{ at } m_k}, 0\}. \tag{8}$$

Moreover, there exists $\tilde{\theta}\in\Theta_I$ such that $\tilde{\theta}_{kk} = \tilde{\theta}^{min}_{kk}$ simultaneously for all k. Thus, when M is fully-ordered and we impose monotonicity, one need not use a linear program to lower bound ν_k or test the sharp null; one can simply plug in the value of $\tilde{\theta}^{min}_{kk}$ to the lower bounds and testable implications given above. ▲

Remark 2 (Identifying Power)**.**
The testable implications we have derived for the sharp null are based on the fact that under the sharp null, there is no effect of the treatment on k-always takers (i.e. $\nu_k = 0$). If the data is compatible with there being no always-takers ($\tilde{\theta}^{min}_{kk} = 0$ for all k), then our lower bounds on the fraction of always-takers affected by the treatment are trivially zero and there is no testable content of the sharp null of full mediation. Intuitively, we therefore have limited testable content when we are "local" to there being no always-takers, i.e. when $\tilde{\theta}^{min}_{kk} \approx 0$ for all k. The expression for $\tilde{\theta}^{min}_{kk}$ in (8) is thus informative about when the testable implications will have bite. In particular, it shows that $\tilde{\theta}^{min}_{kk}$ will tend to be large when there is substantial point mass at $M = m_k$ in the treated group, and when the treatment effect on the survival function of M is small at $M = m_k$. Thus, while our testable implications are valid for any M with a finite number of support points, there will tend to be more identifying power when there is substantial point mass for at least some values of M. ▲

Remark 3 (Binning values of M)**.**
In light of the previous remark, in settings where M is continuous or discrete with many values,

[13] We note that linear programming has been used for tractability in a variety of related but distinct partial identification settings; see, e.g. Mogstad, Torgovitsky and Walters (2024), Ji, Lei and Spector (2024), Yap (2025) for some recent contributions.

it may be tempting to discretize the original M into a small number of bins before applying our results. Let M^{disc} be a discretization of M into K bins (indexed by $k=0,...,K-1$). If we compute the lower bounds for ν_k given in Proposition 3.1 using M^{disc} as the mediator, we obtain valid lower bounds on $P(Y(1,M(1)) \neq Y(0,M(0)) \,|\, M^{disc}(1)=M^{disc}(0)=k)$, i.e. the fraction of people whose outcome depends on treatment status despite having M in the kth bin under both treatments. Our tests for the sharp null thus remain valid if we assume that changes of M within a bin do not affect the outcome, i.e. $Y(d,m)=Y(d,m')$ for all m and m' corresponding to $M^{disc}=k$. This is a strong assumption if taken literally. However, one might reasonably expect that a small change in M should not affect the outcome for most people. This could be captured by the assumption that a change of M within a bin affects no more than ν_{max} fraction of people, so that $P(Y(d,m) \neq Y(d,m') \,|\, M^{disc}(1)=M^{disc}(0)=k) \leqslant \nu_{max}$ for all m,m' in the same bin. In this case, if the sharp null is satisfied, the lower bounds on ν_k given in Proposition 3.1 using M^{disc} as the mediator should all be below ν_{max}. That is, there should exist $\tilde{\theta} \in \Theta_I$ such that $\tilde{\theta}_{kk}\nu_{max} \geqslant \sup_A \Delta_k(A) - \sum_{l:l\neq k} \tilde{\theta}_{lk}$ for all k. ▲

Remark 4 (Functions of the ν_k)**.**

We may sometimes be interested in aggregations of the ν_k across k. For example, the total fraction of always-takers whose outcome is affected by treatment, pooling across k, is given by

$$\bar{\nu} := P(Y(1,M(1)) \neq Y(0,M(0)) \,|\, M(1)=M(0)) = \frac{\sum_k \theta_{kk}\nu_k}{\sum_k \theta_{kk}}.$$

To compute a lower bound on this quantity, we must find $\tilde{\theta}$ and ν to minimize $\frac{\sum_k \tilde{\theta}_{kk}\nu_k}{\sum_k \tilde{\theta}_{kk}}$ subject to the constraints that (5) holds and $\tilde{\theta} \in \Theta_I$. If we reparameterize the problem in terms of $\tilde{\theta}$ and $\tilde{\nu}_k := \tilde{\theta}_{kk}\nu_k$, then both the numerator and denominator of the objective are linear in the parameters, and the constraints are also linear in the parameters if R is a polyhedron. Thus, the problem of minimizing $\frac{\sum_k \theta_{kk}\nu_k}{\sum_k \theta_{kk}}$ over the identified set is a linear-fractional program, which can be recast as a simple linear program via the Charnes and Cooper (1962) transformation. It is thus simple to solve for lower bounds on the total fraction of always-takers affected by treatment, pooling across k. ▲

Remark 5 (Connections to IV testing)**.**

Since testing the sharp null of full mediation is analogous to testing instrument validity—with M playing the role of the endogenous variable and D the instrument—Corollary 3.1 immediately implies sharp testable implications for instrument validity in settings with a binary instrument and multi-valued M.[14,15] The sharp testable restrictions derived here thus may be of independent

[14] Specifically, our results are relevant for testing instrument validity when one assumes the full randomization assumption that the instrument is independent of both potential outcomes and treatments. The implications we derive may not be valid under the weaker notion of independence considered in Kédagni and Mourifié (2020), which imposes only that the instrument is independent of potential outcomes but not potential treatments.

[15] The case where M is multi-dimensional does not have an obvious parallel in the literature on testing instrument validity, since this would correspond to an IV setting with a single instrument but multiple endogenous variables.

interest for the problem of testing instrument validity. Sun (2023) derived non-sharp testable implications of instrument validity in the setting where M is multi-valued but fully-ordered and one imposes monotonicity. His testable restrictions involve only the observable distributions with the minimum and maximum value of M. By contrast, Corollary 3.1 shows that there are in fact testable restrictions coming from all possible values of M, and adding these additional restrictions makes the testable implications sharp. As an illustrative example, suppose that $M \in \{0,1,2\}$ and that we impose monotonicity. Suppose, further, that the treatment D has no impact on the distribution of M. Intuitively, under the sharp null we should then expect the distribution of (Y,M) to be independent of D. Indeed, the sharp testable implication given in Corollary 3.1 corresponds to the restriction that $P(Y \in A, M=k \mid D=1) = P(Y \in A, M=k \mid D=0)$ for all A and $k \in \{0,1,2\}$, which is equivalent to $(Y,M) \perp\!\!\!\perp D$. By contrast, Sun's implications only imply this equality for $k \in \{0,2\}$, which is weaker than full independence. Additionally, while Sun (2023)'s results apply under a monotonicity assumption, our results also imply testable implications under relaxations of monotonicity via a suitable choice of R, as described in Examples 1-4 above. ▲

Remark 6 (Conditional randomization).

Our results in this section assume that D is unconditionally randomly assigned (Assumption 1). If D is conditionally randomly assigned, $D \perp\!\!\!\perp (Y(\cdot,\cdot), M(\cdot)) \mid X$, then in principle all of the testable implications derived above should hold X-almost surely. For example, under the sharp null we should have that $\sup_A \Delta_k(A;X) \leqslant \sum_{l:l \neq k} \theta_{lk}(X)$ (X-a.s.), where $\Delta_k(A;X)$ is the conditional-on-X analog to $\Delta_k(A)$, and $\theta_{lk}(X)$ is the conditional-on-X share of type $G = lk$. If, for example, X takes on a finite number of values, then this amounts to testing the sharp null separately conditional on each value of X. If X is continuous, then practically exploiting the full testable implications of the sharp null will be complicated, however, since it requires estimating the conditional distributions of $(Y,M) \mid X$, and further, the type shares $\theta(X)$ are now a function of X. In Section 5 below, we describe a simpler approach that exploits the testable implications based on the implied marginal distributions of the potential outcomes under conditional unconfoundedness, at the potential loss of sharpness. ▲

Remark 7 (Experiments with missing outcomes).

Li, Sheng and Yu (2025) consider the setting where we have a randomized experiment that generates an outcome $Y^*(D)$. However, Y^* may be missing not-at-random: the observed outcome is $Y = M(D) \cdot Y^*(D)$ where $M(D)$ is an observed indicator for whether the outcome is missing. Li et al. (2025) are interested in the sharp null that $Y^*(1) = Y^*(0)$ (a.s.). Observe that under this sharp null, there is no effect of D on Y for "always-takers" for whom $M(1) = M(0) = k$ for $k = 0,1$. Hence, our tests of the sharp null of full mediation can be used directly to test the sharp null of no treatment effect with missing outcomes in Li et al. (2025). ▲

Remark 8 (Mis-measured mediator.).

Our analysis so far has assumed that the mediator of interest M is observed. In some settings, however, we may only observe a noisy proxy $\tilde{M}$ for M. For example, M could be actual employment and $\tilde{M}$ reported employment on a survey. Assume that $\tilde{M} \perp\!\!\!\perp (Y,D) \mid M$, so that the measurement error is independent of the other variables in the model given the true measurement M. Suppose, further, that the researcher knows the distribution of measurement error, $\tilde{M} \mid M$—for example, the researcher may have access to an auxiliary dataset that contains both survey responses and administrative measures of employment (if not, one could conduct sensitivity analyses to conjectured measurement error distributions). For simplicity, suppose that Y is discrete and M and $\tilde{M}$ have the same support. We then have that $P(Y = y, \tilde{M} = \tilde{m}, D = d) = \sum_m P(Y = y, M = m, D = d) P(\tilde{M} = \tilde{m} \mid M = m)$. We can write this equality as $\tilde{p} = Lp$, where $\tilde{p}$ is the vector collecting probabilities of the form $P(Y = y, \tilde{M} = \tilde{m}, D = d)$ for different $\tilde{m}$; p is analogously the vector collecting the $P(Y = y, M = m, D = d)$ for different m; and L is the $K \times K$ matrix collecting probabilities of the form $P(\tilde{M} = \tilde{m} \mid M = m)$. Provided that L is full-rank, it follows that $p = L^{-1}\tilde{p}$, and thus the distribution of (Y,M,D) is identified. Our results described above, which assume that (Y,M,D) are directly observed, can thus be applied using the implied distributions of (Y,M,D) under this measurement error structure. ▲

3.2 Bounds on average effects for always-takers

So far we have provided lower bounds on ν_k, the fraction of k-always takers who are affected by the treatment despite having $M = m_k$ under both treatments. The ν_k provide a measure of what fraction of always-takers are affected by alternative mechanisms. However, in some settings we may also be interested in the average *magnitude* of the alternative mechanisms for the always-takers. In this section, we derive bounds on

$$ADE_k := E[Y(1,m_k) - Y(0,m_k) \mid G = kk],$$

the average direct effect of the treatment on the outcome for the k-always takers. This provides an alternative measure of the size of the alternative mechanisms for the always-takers.

To derive bounds for ADE_k, we first derive bounds on $E[Y(1,m_k) \mid G = kk]$. Observe that individuals with $M = m_k$ and $D = 1$ must be either k-always takers or lk-compliers. The share of k-always takers among this population is given by $\check{\theta}^1_{kk} := P(G = kk \mid D = 1, M = m_k) = \frac{\theta_{kk}}{P(M = m_k \mid D = 1)}$. It follows that the observable distribution of $Y \mid D = 1, M = m_k$ is a mixture with weight $\check{\theta}^1_{kk}$ on $Y(1,m_k) \mid G = kk$ and weight $(1 - \check{\theta}^1_{kk})$ on the distribution of $Y(1,m_k)$ for lk-compliers. We can thus obtain bounds on $E[Y(1,m_k) \mid G = kk]$ by considering the worst-case scenario where the k-always takers compose the bottom $\check{\theta}^1_{kk}$ fraction of the $Y \mid D = 1, M = m_k$ distribution, and the best-case scenario where they compose the top $\check{\theta}^1_{kk}$ fraction.

The following lemma formalizes this intuition for obtaining bounds on $E[Y(1,m_k)\,|\,G=kk]$, and applies analogous logic to obtain bounds on $E[Y(0,m_k)\,|\,G=kk]$. For ease of notation, we present results in the main text assuming that the distribution of Y is continuous; analogous results without this assumption are given in Lemma A.2 in the Appendix.

Lemma 3.1. *Suppose Assumption 1 holds and that Y is continuously distributed. Let $y^d_q := F^{-1}_{Y|D=d,M=m_k}(q)$ be the qth quantile of $Y\,|\,D=d,M=m_k$. If $\check{\theta}^1_{kk}>0$, then*

$$E[Y\,|\,M=m_k,D=1,Y\leqslant y^1_{\check{\theta}^1_{kk}}]\leqslant E[Y(1,m_k)\,|\,G=kk]\leqslant E[Y\,|\,M=m_k,D=1,Y\geqslant y^1_{1-\check{\theta}^1_{kk}}].$$

Likewise, if $\check{\theta}^0_{kk}:=\frac{\theta_{kk}}{P(M=m_k|D=0)}>0$, then

$$E[Y\,|\,M=m_k,D=0,Y\leqslant y^0_{\check{\theta}^0_{kk}}]\leqslant E[Y(0,m_k)\,|\,G=kk]\leqslant E[Y\,|\,M=m_k,D=0,Y\geqslant y^0_{1-\check{\theta}^0_{kk}}].$$

The bounds are sharp in the sense that there exists a distribution $P^{\dagger}$ for $(Y(\cdot,\cdot),M(\cdot),D)$ consistent with the observable data and Assumption 1 with $\theta_{lk}=P^{\dagger}(G=lk)$ such that the bounds hold with equality.

Lemma 3.1 immediately implies bounds on ADE_k by differencing the inequalities for the expectations of $Y(1,m_k)$ and $Y(0,m_k)$. Note, however, that the bounds in Lemma 3.1 involve the always-taker share $\check{\theta}^d_{kk}=\frac{\theta_{kk}}{P(M=m_k|D=d)}$, which may only be partially identified. It is straightforward to see, however, that the bounds become wider the smaller is $\check{\theta}^d_{kk}$. Intuitively, this is because the most-favorable subdistribution of fraction $\check{\theta}^d_{kk}$ is more favorable the smaller is $\check{\theta}^d_{kk}$, and likewise for the least-favorable subdistribution. Sharp bounds on ADE_k can thus be obtained by plugging $\tilde{\theta}^{min}_{kk}$ into the bounds given in Lemma 3.1, where recall $\tilde{\theta}^{min}_{kk}:=\inf_{\tilde{\theta}\in\Theta_I}\tilde{\theta}_{kk}$ is the minimum value of $\tilde{\theta}_{kk}$ consistent with $\tilde{\theta}\in\Theta_I$. For notation, let $LB_d(\check{\theta}^d_{kk})$ and $UB_d(\check{\theta}^d_{kk})$ denote the lower- and upper-bounds on $E[Y(d,m_k)\,|\,G=kk]$ given in Lemma 3.1 as a function of $\check{\theta}^d_{kk}$. We then have the following bounds on ADE_k.[16]

Proposition 3.2. *Suppose Assumptions 1 and 2 hold and Y is continuously distributed. If $\tilde{\theta}^{min}_{kk}=\inf_{\tilde{\theta}\in\Theta_I}\tilde{\theta}_{kk}>0$, then bounds on ADE_k are given as follows:*

$$LB_1(\check{\theta}^{1,min}_{kk})-UB_0(\check{\theta}^{0,min}_{kk})\leqslant ADE_k\leqslant UB_1(\check{\theta}^{1,min}_{kk})-LB_0(\check{\theta}^{0,min}_{kk})$$

where $\check{\theta}^{d,min}_{kk}:=\frac{\tilde{\theta}^{min}_{kk}}{P(M=m_k|D=d)}$. The lower and upper bounds are sharp in the sense that there exists a distribution $P^{\dagger}$ for $(Y(\cdot,\cdot),M(\cdot),D)$ consistent with the observable data and Assumptions 1 and 2 such that the bound holds with equality.

[16] For settings where Y is not continuous, the analogous result holds if one replaces LB_d and UB_d with the analogous expressions given in Lemma A.2 for the case where Y is not assumed to be continuous.

It is worth noting that in the simple case where M is binary and one imposes monotonicity, the bounds on ADE_k correspond to Lee (2009)'s bounds, where D is viewed as the treatment and M as the "sample selection". In the binary case, the ADE_k can also be viewed as what the statistics literature refers to as principal strata direct effects for the principal strata with $M(1)=M(0)=m_k$ (Frangakis and Rubin, 2002; Zhang and Rubin, 2003).[17] Flores and Flores-Lagunes (2010) observed that such bounds could be used for mediation analysis in the case of binary M—their Proposition 1 matches the bounds given in Lemma 3.1 for the special case where M is binary under monotonicity—although they use this primarily as an intermediate step to derive bounds on the full-population average direct effect of treatment. Our result extends these existing results for the binary case to settings where M may be multi-valued (and where monotonicity may fail).

It is also worth emphasizing that the sharp null of full mediation considered earlier is distinct from the null hypothesis that $ADE_k = 0$ for all k. In particular, the sharp null imposes that the treatment does not have an effect on the outcome for any always-taker, whereas the null that $ADE_k = 0$ imposes that the treatment does not affect the k-always takers on average. This is analogous to the distinction between the sharp null considered by Fisher and the weak null considered by Neyman, applied to the sub-population of always-takers. Thus, we may be able to reject the sharp null in settings where we cannot reject the weaker null that the ADE_k are zero.

4 Inference

The previous section derived testable implications of the sharp null of full mediation, as well as measures of the extent to which it is violated, which involved the distribution of the observable data $(Y,M,D) \sim P_{obs}$. We now derive methods for inference on the sharp null given a sample of N *iid* observations (or clusters) drawn from P_{obs}, $(Y_i,M_i,D_i)_{i=1}^N$. For simplicity of notation, we focus on testing the sharp null, although a simple adaptation of the described approach can be used to test null hypotheses of the form $H_0 : \nu_k \leqslant \nu_k^{ub}\ \forall k$ for any ν_k^{ub} (with the sharp null the special case with $\nu_k^{ub} = 0$ for all k.)

We first comment on the non-standard nature of the inference problem. Recall that the testable implications of the sharp null are equivalent to whether the linear program (7) has a weakly negative solution. However, functions of the observable data enter the constraints of the linear program, and it is well-known that the solution to a linear program can be non-differentiable in the constraints. Second, the function of the observable data in the constraints, $\sup_A \Delta_k(A)$, is itself potentially non-differentiable in the underlying data-generating process. If the outcome Y is discrete, for example,

[17] VanderWeele (2012) argues that one should not interpret the principal stratum effect for compliers as an indirect effect, but rather a combination of the direct and indirect effects (a total effect). This critique does not apply to our analysis of the principal stratum effects for always-takers, since their value of M is unaffected by D, and thus any effects for this subgroup must be direct effects.

then $\sup_A \Delta_k(A) = \sum_y (f_{Y,M=m_k|D=1}(y) - f_{Y,M=m_k|D=0}(y))_+$, where $(x)_+ = \max\{x,0\}$, which is clearly non-differentiable in the partial probability mass functions $f_{Y,M=m_k|D=d}(y) := P(Y=y,M=m_k \,|\, D=d)$ if $f_{Y,M=m_k|D=1}(y) = f_{Y,M=m_k|D=0}(y)$ for any y. Since bootstrap methods are generally invalid when the target parameter is non-differentiable in the underlying data-generating process (Fang and Santos, 2019), we cannot simply bootstrap the solution to (7).

We now show that methods from the moment inequality literature can be used to circumvent these issues. We focus on the case where the distribution of Y is discrete, with support points $y_1,...,y_Q$. As we discuss in Remark 9 below, if Y is continuous, then the tests we derive remain valid if one uses a discretization of Y, although at the potential loss of sharpness. We also focus on the case where R takes the polyhedral form $R = \{\theta \in \Delta : B\theta \leqslant c\}$. To see the connection with moment inequalities, observe that with discrete Y, we have that

$$\sup_A \Delta_k(A) = \sum_{q=1}^{Q} \left(P(Y=y_q,M=m_k \,|\, D=1) - P(Y=y_q,M=m_k \,|\, D=0)\right)_+$$

where again $(x)_+ = \max\{x,0\}$. It follows that the inequality

$$\sup_A \Delta_k(A) \leqslant P(M=m_k \,|\, D=1) - \tilde{\theta}_{kk}$$

holds if and only if there exist $\delta_{k1},...,\delta_{kQ}$ such that

$$\sum_{q=1}^{Q} \delta_{kq} \leqslant P(M=m_k \,|\, D=1) - \tilde{\theta}_{kk} \tag{9}$$

$$\delta_{kq} \geqslant P(Y=y_q,M=m_k \,|\, D=1) - P(Y=y_q,M=m_k \,|\, D=0) \text{ for } q=1,...,Q \tag{10}$$

$$\delta_{kq} \geqslant 0 \text{ for } q=1,...,Q. \tag{11}$$

Hence, the testable implications of the sharp null derived in Corollary 3.1 are equivalent to the statement that there exists some $\tilde{\theta} \in \Theta_I$ and δ such that (9)-(11) hold for all $k=0,...,K-1$.

Observe, further, that $\delta, \tilde{\theta}$ and the observable probabilities enter the inequalities (9)-(11) linearly, and the same is true for the constraints that determine Θ_I. Letting $\omega = (\tilde{\theta}', \delta')'$, it follows that we can write the testable implications of the sharp null as

$$H_0 : \exists \omega \text{ s.t. } C_1\omega - C_2 p \geqslant 0, \tag{12}$$

where C_1, C_2 are known matrices (not depending on the data) and p is a vector that collects probabilities of the forms $P(Y=y_q,M=m_k \,|\, D=d)$ and $P(M=m_k \,|\, D=d)$. A recent literature on moment inequalities has considered testing hypotheses of the above form—in which the nuisance parameter ω enters linearly and with known coefficients C_1—given estimates $\hat{p}$ such that $\sqrt{N}(\hat{p}-p)\to N(0,\Sigma)$ (Andrews et al., 2023; Cox and Shi, 2022; Fang et al., 2023; Cho and Russell, 2024). Under mild conditions, the central limit theorem implies that the vector of conditional sample means $\hat{p}$ is asymptotically normal, and thus existing methods from the aforementioned papers can thus be used directly to test the sharp null of full mediation.

Remark 9 (Discretizing continuous outcomes)**.**
Suppose that the outcome Y is continuously distributed. Let $I_1,...,I_Q$ be disjoint intervals that partition the outcome space, and let Y^{disc} be the discretization of Y that equals j when $Y\in I_j$. Let $\Delta_k^{disc}(A)$ be the analog to $\Delta_k(A)$ using Y^{disc} instead of Y. Observe that

$$\begin{aligned}\sup_A \Delta_k^{disc}(A) &= \sup_A P(Y^{disc}\in A, M=m_k \,|\, D=1) - P(Y^{disc}\in A, M=m_k \,|\, D=0)\\ &= \sup_{A\in\mathcal{A}_{disc}} P(Y\in A, M=m_k \,|\, D=1) - P(Y\in A, M=m_k \,|\, D=0) = \sup_{A\in\mathcal{A}_{disc}} \Delta_k(A)\end{aligned}$$

where $\mathcal{A}_{disc}$ is the σ-algebra generated by $I_1,...,I_Q$. Since $\mathcal{A}_{disc}$ is a subset of the Borel σ-algebra, it follows that $\sup_A \Delta_k^{disc}(A) = \sup_{A\in\mathcal{A}_{disc}} \Delta_k(A) \leqslant \sup_A \Delta_k(A)$. Hence, the testable implications of the sharp null for Y imply the testable implications of the sharp null for any discretization of Y. One can thus obtain valid inference on the sharp null by discretizing the outcome and then using the approach described above with Y^{disc}. Of course, to retain approximate sharpness of the testable implications, one would like to choose a discretization fine enough such that $\sup_A \Delta_k^{disc}(A) \approx \sup_A \Delta_k(A)$. Observe that with a continuous outcome, $\sup_A \Delta_k^{disc}(A) = \sup_A \Delta_k(A)$ if the sign of $f_{Y,M=m_k|D=1}(y) - f_{Y,M=m_k|D=0}(y)$ is constant at all y within the same interval I_j. To obtain approximate sharpness of the testable implications, one would thus like to choose a discretization such that there is a cut-point close to any point where the partial densities cross. A practical tradeoff arises, however, because the validity of the methods described above to test moment inequalities relies on a central limit theorem for the sample probabilities $\hat{p}$, and hence requires that the number of observations per cell (i.e. (Y^{disc},M,D) combination) not be too small. There is thus a trade-off whereby we expect that choosing a smaller number of bins is beneficial for size control but may lead to less sharp testable implications. Matters are further complicated by the fact that the finite-sample power of moment inequality methods may be non-monotonic in the number of bins (as we find in our Monte Carlo simulations below). In Appendix C, we discuss how the choice of bins is closely related to the choice of instrument functions in the literature on conditional moment inequalities (e.g. Andrews and Shi, 2013), which is known to be a challenging problem. Although a formal treatment of the optimal bin choice is beyond the scope of this paper, we provide some practical heuristics following our Monte Carlo simulations in Section 4.1. We note, further, that having a modest number of bins may lead to a more interpretable parameter ν_k. For example, if one uses a discretization using 5 bins based on the quintiles of the outcome, then ν_k corresponds to the fraction of k-always takers whose outcome changes quintile when treated; this may be easier to interpret in some settings than the fraction of always-takers whose outcome is affected at all. ▲

Remark 10 (Incorporating covariates)**.**
As discussed in Remark 6, in some settings we may have access to pre-treatment covariates X, with the treatment conditionally randomly assigned conditional on X, $D \perp\!\!\!\perp (Y(\cdot,\cdot), M(\cdot)) \mid X$. When X is discrete (e.g. an indicator for gender), the testable implications of the sharp null given in Corollary 3.1 hold for each possible value of X. It is straightforward to extend the approach to testing described above to this case: for each possible value of X, we have inequalities of the form $C_1\omega_x - C_2 p_x \geqslant 0$, where ω_x and p_x are analogs to ω and p defined above conditional on $X = x$. We can then use moment inequality methods to test that the inequalities $C_1\omega_x - C_2 p_x \geqslant 0$ hold simultaneously for all values of x. When X is continuous, matters become more complicated because there is now a continuum of inequalities corresponding to each possible value of X, and the nuisance parameter ω_x is an infinite-dimensional function of X. Farbmacher, Guber and Klaassen (2022) and Carr and Kitagawa (2023) develop methods for testing instrument validity using covariates in the setting with a single binary instrument and endogenous treatment, which could be applied in the setting of Section 2 in which M is binary and we impose monotonicity. Extending these approaches to our more general setting, in which the conditional-on-X types shares are only partially-identified, does not appear trivial, and strikes us an interesting question for future research. ▲

4.1 Monte Carlo

To evaluate the methods for inference described above, we conduct Monte Carlo simulations calibrated to our applications to Bursztyn et al. (2020) and Baranov et al. (2020) in Section 6 below. For simplicity, we focus on testing the sharp null under a monotonicity assumption.

Treatment, outcome, and mediator. The treatment, outcome, and mediator in our simulations match those in our empirical applications. For Bursztyn et al. (2020), the treatment is receiving information about other men's beliefs, the outcome is a binary indicator for applying for jobs outside of the home, and the mediator is a binary indicator for job-search service sign-up. For Baranov et al. (2020), the treatment is cognitive behavioral therapy and the outcome is an index of financial empowerment. We consider two mediators, a binary indicator for the presence of a grandmother in the household, and a relationship-quality score, which is a score on a 1-5 scale.

Sample sizes. The sample used for our main analysis of Bursztyn et al. (2020) contains 284 people, with treatment assignment randomized at the individual level (approximately half (139) were treated). For the simulations calibrated to Bursztyn et al. (2020), we draw 284 *iid* observations to match the original sample size. In Baranov et al. (2020), treatment was assigned at the level of a cluster (i.e. at the Union Council level), with a total of 40 clusters (20 treated, 20 control), and a total sample size of approximately 600 individuals (568 or 585 depending on the choice of M). For simulations calibrated to Baranov et al. (2020), we therefore draw 20 independent clusters from each treatment group. Given the small number of clusters, we expect this to be a relatively challenging setting for inference. To evaluate the impact of having a small number of clusters, we also consider alternative simulation designs where we sample 40 or 100 clusters of each treatment type, with a total of 80 and 200 clusters for each design.

Description of DGP. In all of our simulations, we sample the distribution of (Y,M) for control units (or clusters) from the empirical distribution of control units (or clusters) in our applications (i.e. from $(Y,M)|D=0$). For treated units in our simulations, we draw with probability t from the empirical distribution of (Y,M) for treated units, and with probability $1-t$ from the empirical distribution for control units, where $t\in\{0,0.5,1\}$ is a simulation parameter. Thus, when $t=1$, we are sampling both treated and control units in the simulation from the empirical distribution in the data, under which the sharp null is violated. This allows us to assess the power of the various tests. When $t=0$, on the other hand, the distribution of (Y,M) for both treated and control units in the simulation is drawn from the empirical distribution for control units in the original data. This ensures that the testable implications of the sharp null and monotonicity are satisfied, which allows us to evaluate size control. (In fact, the design ensures that all of the implied moment inequalities hold with equality, which is generally a challenging setting for size control for moment inequality methods.) When $t=0.5$, the distribution of (Y,M) for treated units is a mixture of the empirical distribution for treated and control units in the original data. Thus, the sharp null is violated, but the violation is smaller than under the case when $t=1$. Comparing across the cases $t=0.5$ and $t=1$ thereby allows us to evaluate how power changes with the size of the violation of the null.

Methods used. To implement tests based on moment inequalities as described above, we consider the hybrid test proposed by Andrews et al. (2023, henceforth ARP), the conditional conditional chi-squared test proposed by Cox and Shi (2022, henceforth CS),[18] and the test proposed by Fang et al. (2023, henceforth FSST).[19] For comparison to existing methods in the case where M is binary,

[18] More precisely, CS propose a conditional chi-squared test and a "refined" version of this test. Since the refinement is computationally costly with many moments, and only matters when one moment is binding, we only implement the refinement in DGPs with a binary outcome, for which there are fewer moments.

[19] When M is binary, we implement the formulation of the moment inequalities derived in (2) without nuisance parameters. For non-binary M, we use the formulation in (12).

we consider the test for instrument validity proposed by Kitagawa (2015, henceforth K).[20] In the simulations calibrated to Bursztyn et al. (2020), the outcome is binary, and thus no discretization of the outcome is needed. For the simulations calibrated to Baranov et al. (2020), where the outcome takes many values, for the moment inequality methods we consider a discretization of the outcome based on 5 bins in our main specification (see Remark 9). We also consider alternative specifications using 2 and 10 bins. Since the K test does not require a discrete outcome, we use the original continuous outcome when implementing the K test. Implementation of the FSST test requires specifying the moment-selection tuning parameter λ. We consider the two choices recommended by FSST in their Remark 4.2, one of which is data-driven and the other is not. We refer to the resulting tests as FSSTdd and FSSTndd (where 'dd' denotes data-driven). For CS and ARP, we use analytic estimates of the variance of the moments, assuming the data are drawn *iid* in the simulations calibrated to Bursztyn et al. (2020), or that clusters are drawn *iid* in the simulations calibrated to Baranov et al. (2020). Since the K and FSST tests require bootstrap replicates, we use a non-parametric bootstrap at either the individual or cluster level, as appropriate.[21] All tests impose monotonicity as defined in Equation (3).[22] All tests are implemented with nominal size of 5%.

Simulation Results. Table 1 reports the results for simulations designs where we have a binary mediator. This includes the DGP based on Bursztyn et al. (2020) (Panel A), and the DGPs that are based on Baranov et al. (2020) where the considered mediator is the binary indicator for the presence of a grandmother (Panels B-D). Table 2 shows results calibrated to Baranov et al. (2020) using the non-binary relationship quality variable as the mediator. Both tables show the rejection probabilities for each of the methods described above under different simulation designs. To quantify the magnitude of the violations of the sharp null, the table also reports the lower-bound on the fraction of always-takers affected ($\bar{\nu}$).[23]

We first evaluate size control. Recall that DGPs with $t=0$ impose the sharp null of full

[20] For the DGPs based on Baranov et al. (2020), we use a modified version of Kitagawa (2015) to account for clustering.

[21] We have verified that ARP and CS return similar results if we use an analogous bootstrap estimate of the variance rather than the analytic estimates.

[22] As described in the empirical section below, for the multi-valued M in Baranov et al. (2020), the empirical distribution for $M \mid D$ is inconsistent with monotonicity (although the violation is not statistically significant). Our simulation design ensures that the data are consistent with monotonicity under the null DGP ($t=0$). However, the alternative DGPs ($t \in \{0.5,1\}$) are based on the empirical distribution and are therefore inconsistent with monotonicity. Hence, the reported power of tests imposing monotonicity under these alternatives corresponds to their power to jointly detect a violation of the sharp null and a relatively small violation of the monotonicity assumption. We found the ranking of power across methods was identical when we modified the tests to allow for the minimal relaxation of monotonicity consistent with the data, and therefore present the results imposing monotonicity for simplicity and consistency with the other specifications.

[23] For the simulations calibrated to Baranov et al. (2020) with multi-valued M, we compute the lower bound on $\bar{\nu}$ in the same way as described in Footnote 30 in the application section below, which deals with the fact that the empirical distribution shows a small (but statistically insignificant) violation of monotonicity.

Table 1: Simulation results for binary M

Panel A: Bursztyn et al

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.038	0.032	0.030	0.078	0.070
t=0.5	0.036	0.196	0.190	0.116	0.214	0.194
t=1	0.077	0.626	0.632	0.386	0.620	0.584

Panel B: Baranov et al, 40 clusters

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.056	0.154	0.050	0.232	0.212
t=0.5	0.134	0.194	0.206	0.064	0.314	0.270
t=1	0.283	0.570	0.668	0.422	0.750	0.680

Panel C: Baranov et al, 80 clusters

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.044	0.064	0.040	0.132	0.112
t=0.5	0.134	0.322	0.340	0.160	0.410	0.322
t=1	0.283	0.836	0.936	0.846	0.956	0.936

Panel D: Baranov et al, 200 clusters

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.044	0.054	0.030	0.120	0.090
t=0.5	0.134	0.686	0.776	0.618	0.776	0.734
t=1	0.283	0.998	1	1	1	1

Notes: This table contains simulation results for the DGPs where we have a binary mediator. The first column shows the value of t, which determines the distance from the null, as described in the main text. The second column shows the lower-bound on the fraction of always-takers affected by treatment, $\bar{\nu}$. The remaining columns contain the rejection probabilities for each of the methods considered. Panel A shows the results for the DGP based on Bursztyn et al. (2020) and Panels B-D show the results for the DGPs based on Baranov et al. (2020), with the binary grandmother mediator, under different numbers of clusters. In Panels B-D, we use a discretization of the outcome into 5 bins for all tests except the K test. Rejection probabilities are computed over 500 simulation draws, under a 5% nominal significance level.

Table 2: Simulation results for non-binary M

Panel A: Baranov et al, 40 clusters

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.052	0.088	0.274	0.178
t=0.5	0.119	0.066	0.228	0.438	0.374
t=1	0.255	0.166	0.754	0.864	0.828

Panel B: Baranov et al, 80 clusters

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.066	0.048	0.188	0.128
t=0.5	0.119	0.066	0.314	0.582	0.500
t=1	0.255	0.164	0.962	0.994	0.990

Panel C: Baranov et al, 200 clusters

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.046	0.026	0.144	0.108
t=0.5	0.119	0.076	0.542	0.862	0.824
t=1	0.255	0.286	1	1	1

Notes: This table contains simulation results for the DGPs where we have a non-binary mediator. The first column shows the value of t, which determines the distance from the null, as described in the main text. The second column shows the lower-bound on the fraction of always-takers affected by treatment, $\bar{\nu}$. The remaining columns contain the rejection probabilities for each of the inference methods considered. Each panel contains results for the DGPs based on Baranov et al. (2020), where the non-binary relationship-quality mediator is considered, for different numbers of clusters. All tests use a discretization of the outcome based on 5 bins. Rejection probabilities are computed over 500 simulation draws, under a 5% nominal significance level.

mediation. Across nearly all simulation designs, we find that the ARP, CS, and K tests have close to nominal size, with rejection probabilities no larger than 9% for a 5% test. The one notable exception is the simulations in Panel B of Table 1, where there are only 40 independent clusters, in which case CS is somewhat over-sized, with a null rejection probability of 0.15. Doubling the number of clusters to 80 (Panel C) restores approximate size control, however. We find that the FSST tests often have reasonable size control for settings with a large number of independent observations or clusters, but can be substantially over-sized in settings with a small or moderate number of clusters using the two default choices of tuning parameters, particularly with multi-valued M (e.g. rejection probabilities of 0.274 and 0.178 in Table 2, Panel A).

We next evaluate power, focusing on the simulations with $t = 0.5$ and $t = 1$ under which the null is violated. Across all of the simulation designs, the CS test has power similar to or greater than that of ARP. The differences can be substantial in some cases, particularly with multi-valued M (e.g. power of 0.96 vs 0.16 in Panel B of Table 2). Likewise, the power of the FSST tests is similar to or exceeds that of the CS test across nearly all simulation designs, although this comparison must be taken with some caution in cases where the FSST test appears to be over-sized. Finally, we note that in all of the simulations with binary M (Table 1), the power of the three moment inequality tests (ARP, CS, FSST) is either similar to or exceeds that of the K test. This is the case both when the outcome is binary (Panel A), and when the outcome is continuous (Panels B-D). Recall that when the outcome is continuous, the moment inequality tests use a discretization of the outcome to 5 bins, whereas the K test does not use a discretization. The favorable power comparisons in Panels B-D thus suggest that discretization does not come at a large loss of power in this simulation design, although of course this conclusion may be specific to the particular DGP studied here.

Choice of bins. In Appendix Tables 1 and 2, we present results for simulations calibrated to Baranov et al. (2020) using a discretization with 2 or 10 bins, rather than the 5 considered above. The comparisons of size control and power across the methods are similar to the results reported above. However, increasing the number of bins from 5 to 10 exacerbates the size control issues seen with CS in the simulations based on Baranov et al. (2020) with only 40 clusters and binary M, while decreasing the number of bins to 2 improves size control. This is intuitive, since the number of moments used increases with the bin size, and thus we expect the quality of the central limit theorem approximation to be worse with more bins. In Appendix Table 3, we report the median number of independent observations (unique clusters) per (Y^{disc}, M, D) cell in each of the simulation designs. We find that CS exhibits close to nominal coverage in all specifications with 15 or more independent observations per cell, but exhibits moderate size distortions in several (but not all) of the specifications with fewer than 15 observations per cell. In terms of power, we do not find an obvious pattern across bin sizes, with power increasing in the number of bins for some tests/DGPs and decreasing for others. This reflects the fact that although the testable implications become sharper the more bins are used (see Remark 9), the finite-sample power of moment inequality methods often decreases when increasing the number of moments. Based on our simulations, we heuristically recommend that researchers should try to have at least 15 independent observations per cell, acknowledging that there is a potential tradeoff between size control and power. This heuristic roughly aligns with that in Andrews and Shi (2013), who recommend having 10-20 observations per cell in settings with conditional moment inequalities.

Choice of test. Based on our simulations, CS strikes us a reasonable default choice for most empirical settings, given that it has approximate size control across most of our simulation designs and favorable power relative to ARP. However, ARP performs somewhat better in terms of size control in settings with a small number of clusters, and thus may be an attractive alternative for researchers concerned about size control in such settings, albeit at the loss of some power (particularly with multi-valued M). Likewise, FSST may offer power improvements relative to CS in settings with a large number of independent observations, so that size control is not a concern. In our applications below, we report results for CS in the main text; analogous results for ARP and FSST are given in Appendix Table 5.

5 Extension to non-experimental settings

Our results so far have relied on the assumption that the treatment is as good as randomly assigned (Assumption 1). The role of randomization was simply to identify the distribution of potential outcomes and potential mediators under each treatment: randomization ensures that the observable distribution $(Y,M)\,|\,D=d$ corresponds to the distribution of $(Y^{\text{tot}}(d),M(d))$, where $Y^{\text{tot}}(\cdot):=Y(\cdot,M(\cdot))$. In this section, we show that analogous results go through if the distributions of $(Y^{\text{tot}}(d),M(d))$ are identified through some other strategy. One simply substitutes the expressions involving $(Y,M)\,|\,D=d$ in our earlier results with the formulas for $(Y^{\text{tot}}(d),M(d))$ under the alternative identifying assumptions. We first provide a general result extending our results to the case where $(Y^{\text{tot}}(d),M(d))$ is identified, then discuss how this applies to the common settings of instrumental variables, conditional unconfoundedness, and difference-in-differences.

To be more precise, define

$$\Delta_k^*(A):=P(Y^{\text{tot}}(1)\in A,M(1)=m_k)-P(Y^{\text{tot}}(0)\in A,M(0)=m_k)$$

to be the treatment effect on the compound outcome $1\{Y\in A,M=m_k\}$. Note that under randomization, $\Delta_k^*(A)=\Delta_k(A)$. Likewise, define

$$\Theta_I^*:=\{\tilde{\theta}^*\in R\colon \text{ for all } k=0,...,K-1,\ \textstyle\sum_l\tilde{\theta}_{lk}^*=P(M(1)=m_k),\sum_l\tilde{\theta}_{kl}^*=P(M(0)=m_k)\} \tag{13}$$

and observe that $\Theta_I^*=\Theta_I$ under randomization. We then have the following result, analogous to Proposition 3.1, which gives lower bounds on the ν_k in terms of probabilities involving $(Y^{\text{tot}}(d),M(d))$.

Proposition 5.1.

1. *Suppose Assumption 2 holds. Then the true shares θ satisfy*

$$\theta_{kk}\nu_k \geqslant \left(\sup_A \Delta_k^*(A) - \sum_{l:l\neq k}\theta_{lk}\right)_+ \tag{14}$$

for $k=0,...,K-1$. Since $\theta \in \Theta_I^*$, it follows that there exists some $\tilde{\theta}^* \in \Theta_I^*$ such that

$$\tilde{\theta}_{kk}^*\nu_k \geqslant \left(\sup_A \Delta_k^*(A) - \sum_{l:l\neq k}\tilde{\theta}_{lk}^*\right)_+ =: \eta_k \tag{15}$$

for all $k=0,...,K-1$.

2. *The bound in (15) is the sharp bound that only uses information from the marginals $(Y^{tot}(d),M(d))$ for $d=0,1$: if $\tilde{\theta}^* \in \Theta_I^*$, then there exists a distribution P^* for $(Y(\cdot,\cdot),M(\cdot))$ that is consistent with the marginals of $(Y^{tot}(d),M(d))$ for $d=0,1$ such that $P^*(G=lk)=\tilde{\theta}_{lk}^*$ for all l,k, and*

$$\tilde{\theta}_{kk}^* \cdot P^*(Y(1,m_k) \neq Y(0,m_k) \mid G=kk) = \eta_k \tag{16}$$

for all k. Further, $P^*(Y(1,m) \neq Y(0,m) \mid G=lk)=0$ if either $l \neq k$ or $m \notin \{m_l,m_k\}$.

We likewise obtain the following corollary regarding the sharp null of full mediation, analogous to Corollary 3.1.

Corollary 5.1.

1. *Suppose Assumption 2 holds. If the sharp null of full mediation is satisfied, then there exists $\tilde{\theta}^* \in \Theta_I^*$ such that*

$$\sup_A \Delta_k^*(A) \leqslant \sum_{l:l\neq k}\tilde{\theta}_{lk}^* \tag{17}$$

for $k=0,...,K-1$.

2. *The testable implication in (17) is the sharp implication using only the marginals of $(Y^{tot}(d),M(d))$ for $d=0,1$: if there exists a $\tilde{\theta}^* \in \Theta_I^*$ such that (17) holds for all k, then there exists a distribution P^* for $(Y(\cdot,\cdot),M(\cdot))$ consistent with the marginals for $(Y^{tot}(d),M(d))$ and the restriction that $\theta^* \in R$ such that the sharp null of full mediation holds.*

Proposition 5.1 and Corollary 5.1 show that we can bound the fraction of always-takers affected by treatment and test the sharp null so long as the marginal distributions of $(Y^{\text{tot}}(d),M(d))$ are identified. We now outline several settings where these distributions are identified, possibly for some sub-population of interest (e.g. for compliers with respect to an instrument).

Instrumental variables. Suppose rather than D being randomly assigned, we have a valid binary instrument $Z \in \{0,1\}$ for D. For example, in an experiment with imperfect compliance, Z could be the randomized treatment assignment, and D the realized treatment take-up. Suppose that Z satisfies the standard instrument-monotonicity, relevance, exclusion, and independence assumptions (Imbens and Angrist, 1994). To be precise, we assume that $D = D(Z)$, where $D(1) \geqslant D(0)$ (a.s.) and $P(D(1) > D(0)) > 0$, and that $(Y,M) = (Y(D(Z),M(D(Z))),M(D(Z)))$ for $Z \perp\!\!\!\perp (Y(\cdot,\cdot),M(\cdot),D(\cdot))$. These assumptions allow us to identify the LATE of D on Y, i.e. the treatment effect of D on Y for instrument-compliers.[24] They likewise allow us to identify the LATE of D on M. We might then be interested in the extent to which the effect of D on Y for instrument-compliers operates through M. To apply the results in Proposition 5.1 and Corollary 5.1, we need to identify the distributions of $(Y^{\text{tot}}(d),M(d))$ for instrument-compliers. It is well-known, however, that the distributions of potential outcomes for instrument-compliers are non-parametrically identified (see, e.g., Abadie, 2003). In particular, if we define $C^z = 1\{D(1) > D(0)\}$ to be an indicator for being an instrument-complier, then $P(Y^{\text{tot}}(1) \in A, M(1) = m_k \mid C^z = 1)$ is identified as

$$\frac{E[D \cdot 1\{Y \in A, M = m_k\} \mid Z = 1] - E[D \cdot 1\{Y \in A, M = m_k\} \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}. \tag{18}$$

In words, the probability that $Y^{\text{tot}}(1) \in A$ and $M(1) = m_k$ for instrument-compliers corresponds to the population IV estimand using the compound outcome $D \cdot 1\{Y \in A, M = m_k\}$. We can likewise identify $P(Y^{\text{tot}}(0) \in A, M(0) = m_k)$ using the population IV estimand with the compound outcome $-(1-D) \cdot 1\{Y \in A, M = m_k\}$.[25]

Conditional unconfoundedness. Suppose that D is as good as randomly assigned conditional on observable characteristics, $D \perp\!\!\!\perp (Y(\cdot,\cdot),M(\cdot)) \mid X$. Under the overlap condition that $\eta < E[D \mid X] < 1 - \eta$ for some $\eta > 0$, the distributions of $(Y^{\text{tot}}(d),M(d))$ are non-parametrically identified by re-weighting the observed outcomes by the propensity score $p(X) := E[D \mid X]$,

$$\begin{aligned} P(Y^{\text{tot}}(1) \in A, M(1) = m_k) &= E\left[\frac{D}{p(X)} 1\{Y \in A, M = m_k\}\right] \\ P(Y^{\text{tot}}(0) \in A, M(0) = m_k) &= E\left[\frac{1-D}{1-p(X)} 1\{Y \in A, M = m_k\}\right]. \end{aligned} \tag{19}$$

[24] We use the phrase "instrument-compliers" to refer to compliers with respect to the instrument, i.e. with $D(1) > D(0)$, in contrast to the use of "compliers" elsewhere in the paper, which refers to individuals with $M(1) \neq M(0)$.

[25] Note that the instrument exclusion restriction implies that Z affects Y only through D, and the sharp null implies that D affects Y only through M. Hence, under the sharp null, Z affects Y only through M. A simple approach to testing the sharp null in IV settings is thus to apply the results in Corollary 3.1 for experiments, relabeling the randomized treatment as Z and ignoring the endogenous take-up D. In Section B.3 we show that this is equivalent to applying Corollary 5.1 when one imposes monotonicity of $M(d)$ in d, but otherwise does not exhaust all of the information in the data.

Hence, one can apply the results in Proposition 5.1 and Corollary 5.1 to obtain lower-bounds on the fraction of always-takers affected by the treatment, and to test the sharp null.[26] ▲

Difference-in-differences. Consider a two-period setting where no units are treated in the first period and units with $D = 1$ are treated in the second period. Under a parallel trends assumption for $Y^{\text{tot}}(0)$, we can identify $E[Y_2^{\text{tot}}(1) - Y_2^{\text{tot}}(0) \mid D = 1]$, the average treatment effect on the treated (ATT) in period 2 (where the 2 subscript denotes the second time period). Likewise, under a parallel trends assumption for $M(0)$, we can identify $E[M_2(1) - M_2(0) \mid D = 1]$, the ATT on M_2. We may then be interested in the extent to which the ATT for Y_2 is driven by the effect of the treatment on M_2. However, the standard parallel trends assumptions for $Y^{\text{tot}}(0)$ and $M(0)$ identify only the counterfactual *means* for the treated group and not the counterfactual *distributions*, as would be required to apply the results in Proposition 5.1 and Corollary 5.1. However, several papers have developed extensions of the standard difference-in-differences approach that allow one to infer the full counterfactual distribution for the treated group (Athey and Imbens, 2006; Callaway and Li, 2019; Roth and Sant'Anna, 2023). These approaches could be applied to identify the distributions of $(Y_2^{\text{tot}}(d), M_2(d)) \mid D = 1$, which could then be used in conjunction with Proposition 5.1 and Corollary 5.1 to examine the extent to which the ATT on Y_2 is driven by the effect on M_2. ▲

The approach to inference described in Section 4 naturally extends to these settings as well. In Section 4, we considered inference based on a vector of estimates $\hat{p}$, where each element of $\hat{p}$ corresponded to an estimate of a probability of the form $P(Y^{\text{tot}}(d) \in A, M = m_k)$ under the assumption of randomly assigned treatment. To test the sharp null under the identifying strategies described above, one simply replaces the $\hat{p}$ in Section 4 with analogous estimates of $P(Y^{\text{tot}}(d) \in A, M = m_k)$ derived under the alternative identifying assumptions. For example, in IV settings we could use two-stage least squares estimates based on the sample analog to the identification result in (18); under conditional unconfoundedness, we could use inverse-probability weighting estimates based on sample analogs to equation (19) (one could likewise use outcome-modeling or doubly-robust methods); and in difference-in-differences settings we could use estimates obtained from any of the distributional difference-in-differences approaches.

[26] We note that the approach just described uses only the information about the implied marginal distributions of $(Y^{\text{tot}}(d), M(d))$. As described in Remarks 6 and 10, under conditional unconfoundedness one could potentially use information on the conditional distributions $(Y^{\text{tot}}(d), M(d)) \mid X$. Doing so in practice is complicated, however, since it requires estimating conditional distributions and involves the infinite-dimensional conditional type shares, $\theta(X)$. The approach just-described is simpler, since it does not involve estimating full conditional distributions and has a finite-dimensional parameter θ. However, it may be conservative since it does not exploit all the information available.

6 Empirical applications

6.1 Bursztyn et al. (2020) revisited

We now revisit our application to Bursztyn et al. (2020) from Section 2. Recall that our treatment D is random assignment to an information treatment about other men's beliefs about women working outside the home, M is sign-up for the job-search service, and Y is an indicator for whether the wife applies for jobs outside of the home. For our main specification, we restrict attention to the majority of men who at baseline under-estimate other men's beliefs, so that the monotonicity assumption that treatment weakly increases job-search service is plausible. (We find similar results when including all men; see Appendix D.)

Statistical significance. Recall from Figure 1 that the testable implications of the sharp null were rejected based on the empirical distribution. Using the approach to inference described above, we find these violations are in fact statistically significant, with a p-value of 0.02 using the CS test.[27] (We obtain similar results using the other tests; see Appendix Table 5.) The data thus provides strong evidence that the impact of the information treatment on long-run labor market outcomes does not operate solely through the sign-up for the job-search service. In particular, there are some never-takers who would not sign up for the service under either treatment who are nevertheless induced to apply for jobs by the treatment. We thus see that, for at least some people, the information treatment has meaningful impact outside of the lab, beyond its impact on job-search service sign-up.

Magnitudes of alternative mechanisms. How large are the effects of the information treatment for those who are not induced to sign-up for the job-search service? Proposition 3.1 gives us a lower bound on the fraction of the always-takers/never-takers who are affected by the treatment despite having no effect on job-search service signup. Our estimates of the lower bounds suggest that at least 11 percent of "never-takers" who would not be signed up for the job-search service under either treatment are nevertheless affected by the treatment. (We obtain a trivial lower-bound of 0 for the "always-takers".) Applying the results in Proposition 3.2, we also estimate lower and upper bounds on the average effect for these never-takers of 0.11 to 0.18.[28] For comparison, our estimate of the overall average treatment effect is 0.12. The effect for never-takers is thus of a fairly similar magnitude to that of the total population, despite the fact that they have no change in job-search service signup. If we were willing to assume that the direct effects (i.e. effects not through the job-search service)

[27] Since the outcome is binary, no discretization is needed for this application. The p-value reported here is the smallest value of α for which the test rejects.

[28] Because the outcome is binary, the lower bound for the average effect corresponds exactly to our lower bound on the fraction of always-takers affected.

were similar between always-takers, never-takers, and compliers (granted, a strong assumption), this would imply that the majority of the total effect operates through the information treatment.

Robustness to monotonicity violations. Our baseline results impose the monotonicity assumption that receiving the information that other men are more open to women working than one initially thought only increases job-search service sign-up. This could be violated if, for example, there is measurement error in the initial elicitation of beliefs, so that some men included in our sample actually initially over-estimated other men's beliefs. To explore robustness to violations of the monotonicity assumption, we re-compute our bounds on the fraction of never-takers affected allowing for up to $\bar{d}$ fraction of the population to be defiers. We find that the estimated lower-bound is positive for $\bar{d}$ up to 0.07, which corresponds to 7% of the population being defiers, or put otherwise, 0.33 defiers for every complier.

6.2 Baranov et al. (2020)

We next examine the setting of Baranov et al. (2020). They present long-run results on an RCT that randomized access to a cognitive behavioral therapy (CBT) program intended to reduce depression for pregnant women and recent mothers. In a seven-year followup, they find that the program substantially reduced depression and increased measures of women's financial empowerment, such as having control over finances and working outside of the home. They are then interested in the mechanisms by which treating depression increases financial empowerment. They therefore examine a variety of intermediate outcomes. Two of the outcomes for which they find positive effects of the treatment are the presence of a grandmother in the household (a proxy for family support) and the women's self-reported relationship quality with the husband (on a 1-5 scale). They write (p. 849):

> These results suggest that improved social support within the household, either through a relationship with the husband or asking grandmothers for help, might be a mechanism underlying the effectiveness of this CBT intervention.

The tools developed above allow us to test the completeness of these conjectured mechanisms. Can the presence of a grandmother or improved relationship quality, either individually or together, explain the impact on financial empowerment, or must there be other mechanisms at play as well? We begin by analyzing each of these mechanisms separately, and then turn to studying the combination of the two.

Since the outcome in Baranov et al. (2020) is a continuous index, we rely on a discretization using 5 bins, which we found in our Monte Carlo simulations calibrated to this application led to a reasonable tradeoff between size control and power (although with moderate size distortions for the setting with binary M). This also roughly aligns with our heuristic for choosing the bin size in the setting with binary M, as it yields a median cell count of 14. In the setting with non-binary M, this leads to a median cell count of 8, which is somewhat below our heuristic threshold of 15; using 2 bins would deliver a count of 15. Nevertheless, in the main text we present results using 5 bins to maintain consistency in the outcome variable across different specifications, and because the Monte Carlo results suggest that this choice performs well in this application with multi-valued M despite the small cell size. This also gives the ν_k naturally-interpretable units as the fraction of always-takers whose outcome changes quintile when treated. In Appendix Table 4, we find qualitatively similar results using 2 or 10 bins.

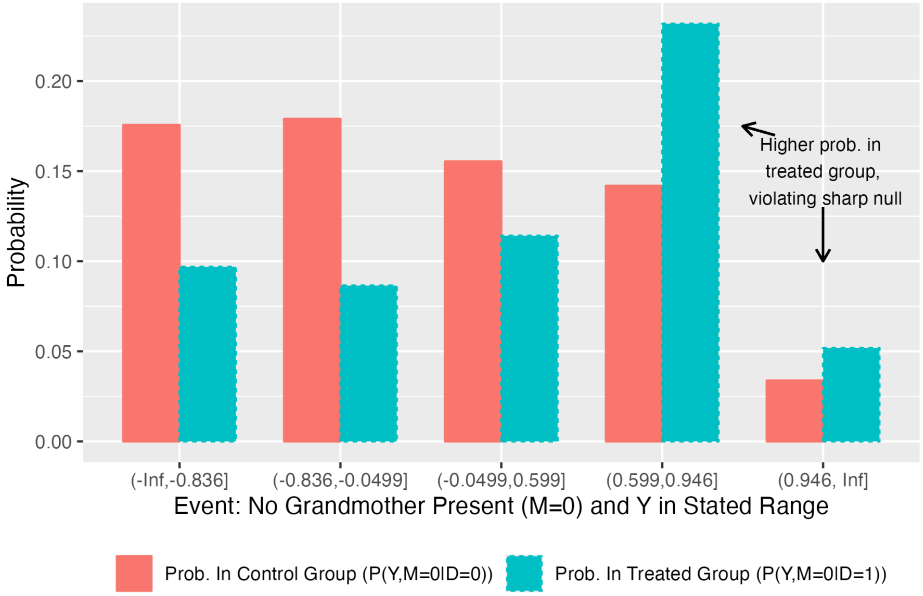
Grandmother Mechanism. We first examine whether the effects of the intervention can be explained through the binary mechanism of whether a grandmother is present in the household (measured at the 7-year follow-up). Figure 3 shows estimates of $P(Y=y, M=0 \mid D=d)$ for both $d=1$ and $d=0$, similar to Figure 1 for our previous application. If one imposes monotonicity, then as derived in Section 2 we should have that $P(Y=y, M=0 \mid D=1) \leqslant P(Y=y, M=0 \mid D=0)$ for all values of y. As shown in the figure, however, this inequality appears to be violated at large values of y, suggesting that the outcome for some never-takers improved when receiving the treatment. These violations of the sharp null are statistically significant (CS $p = 0.02$). Our estimates of the lower bound derived in Proposition 3.1 imply that at least 19 percent of never-takers are affected by the treatment. Thus, we can reject that the entirety of the treatment effect operates through increased grandmother presence in the home.[29] These conclusions rely on the monotonicity assumption that receiving CBT weakly increases the presence of the grandmother; this could be violated if, for example, some grandmothers were present when the mother was struggling but decided they were no longer needed as the mother improved. As before, we can explore robustness to allowing for defiers: our estimated lower bounds on the fraction of never-takers affected remain positive unless we allow for at least 11 percent of the population to be defiers, or equivalently, 0.51 defiers per complier.

Relationship quality mechanism. We next examine relationship quality (as of the 7-year follow-up) as the mechanism, which is measured on a 1-5 scale. We can thus apply the methods for multi-valued M developed in Section 3. Under the monotonicity assumption that CBT improves the relationship with the husband, we obtain a point estimate of the lower bound on the fraction of always-takers affected (pooling across different values of M) of 10%, and we reject the sharp null using CS ($p = 0.03$).[30] There is thus some evidence that the effect of CBT on financial empowerment

[29] Specifically, we can reject that the effect operates through increasing *long-run* grandmother presence, as measured at the 7-year follow-up. The results are less conclusive using the presence of a grandmother at the 1-year follow-up: we obtain $p=0.19$, although the point estimates suggest that 14 percent of never-takers are affected by treatment.

[30] The monotonicity assumption requires that the population CDF of $M \mid D=1$ is everywhere smaller than the population CDF of $M \mid D=0$. This is satisfied at three of the four support points of the empirical distribution. However, the empirical CDF in the treated group is 0.015 larger at $M = 4$, although this difference is not

Figure 3: Testable Implications of the Sharp Null for the Grandmother Mediator in Baranov et al. (2020)



Note: This figure shows testable implications of the sharp null of full mediation in Baranov et al. (2020), similar to Figure 1. The mediator is presence of a grandmother in the home. The bars show estimates of probabilities of the form $P(Y^{disc}=y, M=0 \mid D=d)$, where Y^{disc} is a discretization of the outcome (an index of mother's financial empowerment) into 5 bins. Under the sharp null of full mediation, we should have that $P(Y^{disc}=y, M=0 \mid D=0) \geqslant P(Y^{disc}=y, M=0 \mid D=1)$, but this appears to be violated for large values of y, as indicated with the black arrows.

does not operate entirely through improvements in relationship quality. The lower bound on the fraction of always-takers affected remains positive allowing for up to 8% of the population to be defiers.

Combinations of mechanisms. Can the combination of the grandmother and relationship-quality mechanisms explain the improvement in financial empowerment? To evaluate this, we consider the case where M is a vector containing both candidate mechanisms. If we impose the monotonicity assumption that treatment increases each of the elements of M, we obtain an estimated lower bound on the fraction of always-takers affected of 7%. However, this is not statistically significant at conventional levels (CS $p=$ 0.65). Although the point estimates suggest some violations, we thus do not significantly reject the null hypothesis that the combination of these two mechanisms, which the authors interpret broadly as proxies for "social support within the household", can explain the effect of CBT on financial empowerment. This of course does not

statistically significant from zero (p=0.75). Thus, the empirical distribution violates monotonicity, although we cannot reject that monotonicity holds in the population. To compute our estimate of the lower bound on the fraction of always-takers affected using the empirical distribution, we therefore allow for the minimum number of defiers compatible with the empirical distribution of $M \mid D$ (0.015). We apply an analogous approach when considering the grandmother and relationship-quality mechanisms jointly.

establish that no other mechanisms are at play, but rather that the data are statistically consistent with this null hypothesis at conventional levels.

7 Conclusion

This paper develops tests for the "sharp null of full mediation" that the effect of a treatment D on an outcome Y operates only through a conjectured set of mediators M. A key observation is that when M is binary, existing tools for testing the validity of the LATE assumptions can be used for testing the null. We develop sharp testable implications in a more general setting that allows for multi-valued and multi-dimensional M, and allows for relaxations of the monotonicity assumption. Our results also provide lower bounds on the size of the alternative mechanisms for always-takers. We illustrate the usefulness of these tests in two empirical applications.

Future work might extend the analysis in this paper in several directions. First, our analysis focuses on the case where M is discrete. Although one can discretize M under the assumptions described in Remark 3, an interesting question for future work is whether one can impose alternative assumptions that allow for testing the sharp null directly when M is continuous. One potentially fruitful direction is to explore whether methods for testing instrument validity with a continuous treatment (e.g. D'Haultfœuille, Hoderlein and Sasaki, 2024) can be adapted to this setting. Second, our current analysis allows the potential outcomes to depend arbitrarily on M, and does not impose any assumptions on how M is assigned. In some settings, however, it may be reasonable to restrict the magnitude of the effect of M on Y, or to restrict the degree of endogeneity of M. Incorporating such restrictions may lead to sharper testable implications. Finally, it may be interesting to extend our results to settings with non-binary treatments.

References

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, April 2003, *113* (2), 231–263.

Andrews, Donald WK and Xiaoxia Shi, "Inference based on conditional moment inequalities," *Econometrica*, 2013, *81* (2), 609–666.

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes, "Inference for Linear Conditional Moment Inequalities," *The Review of Economic Studies*, January 2023, p. rdad004.

Angrist, Joshua D. and Peter Hull, "Instrumental variables methods reconcile intention-to-screen effects across pragmatic cancer screening trials," *Proceedings of the National Academy* *of Sciences*, December 2023, *120* (51), e2311556120. Publisher: Proceedings of the National Academy of Sciences.

_ , Parag A. Pathak, and Roman A. Zarate, "Choice and consequence: Assessing mismatch at Chicago exam schools," *Journal of Public Economics*, July 2023, *223*, 104892.

Armstrong, Timothy B., "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, August 2014, *181* (2), 92–116.

Athey, Susan and Guido W. Imbens, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

_ , Raj Chetty, Guido Imbens, and Hyunseung Kang, "Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index," August 2024. arXiv:1603.09326 [stat].

Balke, Alexander and Judea Pearl, "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, September 1997, *92* (439), 1171–1176. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.1997.10474074.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko, "Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial," *American Economic Review*, March 2020, *110* (3), 824–859.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott, "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review*, October 2020, *110* (10), 2997–3029.

Callaway, Brantly and Tong Li, "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 2019, *10* (4), 1579–1618.

Carr, Thomas and Toru Kitagawa, "Testing Instrument Validity with Covariates," September 2023. arXiv:2112.08092 [econ].

Charnes, A. and W. W. Cooper, "Programming with linear fractional functionals," *Naval Research Logistics Quarterly*, 1962, *9* (3-4), 181–186.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen, "Intersection Bounds: Estimation and Inference," *Econometrica*, 2013, *81* (2), 667–737. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA8718.

Chetverikov, Denis, "Adaptive Tests of Conditional Moment Inequalities," *Econometric Theory*, 2018, *34* (1), 186–227. Publisher: Cambridge University Press.

Cho, JoonHwan and Thomas M. Russell, "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments," *Journal of Business & Economic Statistics*, April 2024, *42* (2), 563–578.

Cox, Gregory and Xiaoxia Shi, "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," *The Review of Economic Studies*, March 2022, p. rdac015.

Deuchert, Eva, Martin Huber, and Mark Schelker, "Direct and Indirect Effects Based on Difference-in-Differences With an Application to Political Preferences Following the Vietnam Draft Lottery," *Journal of Business & Economic Statistics*, October 2019, *37* (4), 710–720.

D'Haultfœuille, Xavier, Stefan Hoderlein, and Yuya Sasaki, "Testing and relaxing the exclusion restriction in the control function approach," *Journal of Econometrics*, March 2024, *240* (2), 105075.

Fang, Zheng and Andres Santos, "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, January 2019, *86* (1), 377–412.

_ , _ , Azeem M. Shaikh, and Alexander Torgovitsky, "Inference for Large-Scale Linear Systems With Known Coefficients," *Econometrica*, 2023, *91* (1), 299–327. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18979.

Farbmacher, Helmut, Raphael Guber, and Sven Klaassen, "Instrument Validity Tests With Causal Forests," *Journal of Business & Economic Statistics*, April 2022, *40* (2), 605–614.

Flores, Carlos and Alfonso Flores-Lagunes, "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," Working paper January 2010.

Frangakis, Constantine E. and Donald B. Rubin, "Principal Stratification in Causal Inference," *Biometrics*, March 2002, *58* (1), 21–29.

Frölich, Markus and Martin Huber, "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, November 2017, *79* (5), 1645–1666.

Huber, Martin, "A review of causal mediation analysis for assessing direct and indirect treatment effects," Technical Report 2019.

_ and Giovanni Mellace, "Testing Instrument Validity for Late Identification Based on Inequality Moment Constraints," *The Review of Economics and Statistics*, 2015, *97* (2), 398–411. Publisher: The MIT Press.

_ , Lukas Laffers, and Giovanni Mellace, "Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations Under Endogeneity and Noncompliance," *Journal of Applied Econometrics*, 2017, *32* (1), 56–79.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, February 2010, *25* (1), 51–71. Publisher: Institute of Mathematical Statistics.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475. Publisher: [Wiley, Econometric Society].

Ji, Wenlong, Lihua Lei, and Asher Spector, "Model-Agnostic Covariate-Assisted Inference on Partially Identified Causal Effects," November 2024. arXiv:2310.08115 [econ].

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

Kédagni, Désiré and Ismael Mourifié, "Generalized instrumental inequalities: testing the instrumental variable independence assumption," *Biometrika*, September 2020, *107* (3), 661–675.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, July 2009, *76* (3), 1071–1102.

Li, Xinran, Peizan Sheng, and Zeyang Yu, "Randomization Inference with Sample Attrition," July 2025. arXiv:2507.00795 [econ].

Ludwig, Jens, Jeffrey R Kling, and Sendhil Mullainathan, "Mechanism Experiments and Policy Evaluations," *Journal of Economic Perspectives*, August 2011, *25* (3), 17–38.

Miles, Caleb H, "On the causal interpretation of randomised interventional indirect effects," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, September 2023, *85* (4), 1154–1172.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters, "Policy evaluation with multiple instrumental variables," *Journal of Econometrics*, July 2024, *243* (1), 105718.

Mourifié, Ismael and Yuanyuan Wan, "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, May 2017, *99* (2), 305–313.

Pearl, Judea, "Direct and indirect effects," in "Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence" San Francisco 2001, pp. 411–420.

Robins, James M. and Sander Greenland, "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 1992, *3* (2), 143–155. Publisher: Lippincott Williams & Wilkins.

Roth, Jonathan and Pedro HC Sant'Anna, "When is parallel trends sensitive to functional form?," *Econometrica*, 2023, *91* (2), 737–747. Publisher: Wiley Online Library.

Schenk, Timo, "Mediation Analysis in Difference-in-Differences Designs," Technical Report 2023.

Sun, Zhenting, "Instrument validity for heterogeneous causal effects," *Journal of Econometrics*, 2023, *237* (2), 105523.

VanderWeele, Tyler J., "Comments: Should Principal Stratification Be Used to Study Mediational Processes?," *Journal of Research on Educational Effectiveness*, July 2012, *5* (3), 245–249.

_ , "Mediation Analysis: A Practitioner's Guide," *Annual Review of Public Health*, 2016, *37*, 17–32.

Villani, Cédric, *Optimal Transport*, Vol. 338 of *Grundlehren der mathematischen Wissenschaften*, Berlin, Heidelberg: Springer, 2009.

Wang, Linbo, James M. Robins, and Thomas S. Richardson, "On falsification of the binary instrumental variable model," *Biometrika*, March 2017, *104* (1), 229–236.

Yap, Luther, "Sensitivity of Policy Relevant Treatment Parameters to Violations of Monotonicity," *Working paper*, 2025.

Zhang, Junni L. and Donald B. Rubin, "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death"," *Journal of Educational and Behavioral Statistics*, December 2003, *28* (4), 353–368. Publisher: American Educational Research Association.

For Online Publication

A Proofs of Results in Main Text

We prove Proposition 5.1 before Proposition 3.1, since Proposition 3.1 follows as a corollary to Proposition 5.1. In all proofs presented in Appendix A, we relabel m_k as k for notational convenience.

Proof of Proposition 5.1

Proof of Part 1. By the law of total probability, we have that

$$\begin{aligned} P(Y^{\text{tot}}(1)\in A, M(1)=k) &= \sum_l P(M(0)=l, M(1)=k)\cdot P(Y(1,k)\in A \mid M(0)=l, M(1)=k) \\ &= \sum_l \theta_{lk}\cdot P(Y(1,k)\in A \mid G=lk) \end{aligned}$$

and similarly,

$$P(Y^{\text{tot}}(0)\in A, M(0)=k) = \sum_l \theta_{kl} P(Y(0,k)\in A \mid G=kl).$$

Combining the previous two displays, it follows that

$$\begin{aligned} &\theta_{kk}(P(Y(1,k)\in A \mid G=kk) - P(Y(0,k)\in A \mid G=kk)) \\ &= P(Y^{\text{tot}}(1)\in A, M(1)=k) - P(Y^{\text{tot}}(0)\in A, M(0)=k) - \\ &\qquad \sum_{l:l\neq k}\theta_{lk} P(Y(1,k)\in A \mid G=lk) + \sum_{l:l\neq k}\theta_{kl} P(Y(0,k)\in A \mid G=kl) \\ &\geqslant P(Y^{\text{tot}}(1)\in A, M(1)=k) - P(Y^{\text{tot}}(0)\in A, M(0)=k) - \sum_{l:l\neq k}\theta_{lk} \end{aligned}$$

where the inequality uses the fact that probabilities are bounded between 0 and 1. Taking a sup over Borel sets A, we obtain that

$$\begin{aligned} \theta_{kk} TV_{kk} &\geqslant \sup_A \{P(Y^{\text{tot}}(1)\in A, M(1)=k) - P(Y^{\text{tot}}(0)\in A, M(0)=k)\} - \sum_{l:l\neq k}\theta_{lk} \\ &= \sup_A \Delta_k^*(A) - \sum_{l:l\neq k}\theta_{lk}, \end{aligned}$$

where

$$TV_{kk} := \sup_A \{P(Y(1,k)\in A \mid G=kk) - P(Y(0,k)\in A \mid G=kk)\}$$

is the total variation (TV) distance between $Y(1,k) \mid G=kk$ and $Y(0,k) \mid G=kk$. To establish the first claim, it thus suffices to show that $P(Y(1,k) \neq Y(0,k) \mid G = kk) \geqslant TV_{kk}$. Recall, however, that the TV distance is the Wasserstein-0 distance (see, e.g., Villani, 2009), and thus

$$TV_{kk} = \inf_{\substack{Q \text{ s.t.} \\ (Y(1,k),Y(0,k)) \sim Q \\ Y(1,k) \sim P_{Y(1,k)|G=kk} \\ Y(0,k) \sim P_{Y(0,k)|G=kk}}} E_Q[1\{Y(1,k) \neq Y(0,k)\}],$$

where $P_{Y(d,k)|G=kk}$ is the marginal distribution of $Y(d,k) \mid G = kk$. Since $E_Q[1\{Y(1,k) \neq Y(0,k)\}] = P_Q(Y(1,k) \neq Y(0,k))$, it follows (from the definition of the inf) that $P(Y(1,k) \neq Y(0,k) \mid G = kk) \geqslant TV_{kk}$, which completes the proof of the first claim. The second claim is immediate from the fact that $\theta \in \Theta_I^*$ by construction.

Proof of Part 2. Fix $\tilde{\theta}^* \in \Theta_I^*$. Since $M(d)$ has finite support, there exists a dominating, positive σ-finite measure μ[31] and densities $f_{Y^{\text{tot}}(d)|M(d)}$ such that for $d = 0,1$ and all k,

$$P(Y^{\text{tot}}(d) \in A, M(d) = k) = P(M(d) = k) \cdot \int_A f_{Y^{\text{tot}}(d)|M(d)=k} d\mu.$$

We define the partial density $f_{Y^{\text{tot}}(d),M(d)=k}(y) := P(M(d) = k) \cdot f_{Y^{\text{tot}}(d)|M(d)=k}(y)$. Note that

$$\sup_A \Delta_k^*(A) = \int_{\mathcal{Y}} \left(f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k}\right)_+ d\mu.$$

We begin by proving the following lemma.

Lemma A.1. *Suppose that there exist collections of probability densities (measurable wrt μ), $\mathcal{F}_1 := (f^*_{Y(1,k)|G=lk})_{l=0}^{K-1}$ and $\mathcal{F}_0 := (f^*_{Y(0,k)|G=kl})_{l=0}^{K-1}$, such that for all k,*

$$f_{Y^{tot}(1),M(1)=k} = \sum_l \tilde{\theta}^*_{lk} f^*_{Y(1,k)|G=lk} \tag{20}$$

$$f_{Y^{tot}(0),M(0)=k} = \sum_l \tilde{\theta}^*_{kl} f^*_{Y(0,k)|G=kl}, \tag{21}$$

and for all k with $\tilde{\theta}^*_{kk} > 0$,

$$\tilde{\theta}^*_{kk} TV_{kk} = \eta_k, \text{ where } TV_{kk} := \int_{\mathcal{Y}} \left(f^*_{Y(1,k)|G=kk} - f^*_{Y(0,k)|G=kk}\right)_+ d\mu. \tag{22}$$

[31] Specifically, we can take $\mu(\cdot) = \sum_{m,d} \mu_{Y^{\text{tot}}(d)|M(d)=m}(\cdot)$, where $\mu_{Y^{\text{tot}}(d)|M(d)=m}$ is the probability measure for $Y^{\text{tot}}(d) \mid M(d) = m$ if $P(M(d) = m) > 0$ and zero otherwise. By construction μ is a positive σ-finite dominating measure, and hence the densities exist by the Radon-Nikodym theorem for m,d such that $P(M(d) = m) > 0$. For m,d such that $P(M(d) = m) = 0$, we can trivially set $f_{Y^{\text{tot}}(d)|M(d)=m}$ to be any probability density wrt μ.

Then there exists a joint distribution P^* for $(Y(\cdot,\cdot),M(\cdot))$ satisfying the conditions of Proposition 5.1 part 2, i.e. such that $P^*(G=lk)=\tilde{\theta}^*_{lk}$ for all l,k; $\tilde{\theta}^*_{kk}P^*(Y(1,k)\neq Y(0,k)\,|\,G=kk)=\eta_k$ for all k; and $P^*(Y(1,m)\neq Y(0,m)\,|\,G=lk)=0$ if either $l\neq k$ or $m\notin\{l,k\}$.

Proof of Lemma. We will construct a distribution P^* under which $(Y(1,k),Y(0,k))\perp\!\!\!\perp(Y(1,k'),Y(0,k'))\,|\,G$ for $k\neq k'$. That is, we will construct a P^* that takes the form

$$P^*((Y(0,0),Y(1,0))\in B_0,\dots,(Y(0,K-1),Y(1,K-1))\in B_{K-1},G=lk)=\tilde{\theta}^*_{lk}\cdot\prod_{m=0}^{K-1}\int_{B_m}f^*_{(Y(1,m),Y(0,m))|G=lk}\,d\tilde{\mu}_{m,lk},$$

where the $\tilde{\mu}_{m,lk}$ are measures on $\mathcal{Y}^2$ with one-dimensional marginals dominated by μ. Note that this construction implies that $P^*(G=lk)=\tilde{\theta}^*_{lk}$. Let $f^{**}_{Y(d,k)|G=g}$ denote the implied marginal distribution over $Y(d,k)\,|\,G=g$ under P^*. Then P^* matches the marginals of $(Y^{\text{tot}}(d),M(d))$ if and only if

$$P^*(M(1)=k)=P(M(1)=k)\text{ for all }k \tag{23}$$

$$P^*(M(0)=k)=P(M(0)=k)\text{ for all }k \tag{24}$$

$$\sum_l\tilde{\theta}^*_{lk}f^{**}_{Y(1,k)|G=lk}=f_{Y^{\text{tot}}(1),M(1)=k}\;(\mu\text{-a.s., for all }k) \tag{25}$$

$$\sum_l\tilde{\theta}^*_{kl}f^{**}_{Y(0,k)|G=kl}=f_{Y^{\text{tot}}(0),M(0)=k}\;(\mu\text{-a.s., for all }k) \tag{26}$$

Note that by construction, $P^*(M(1)=k)=\sum_l\tilde{\theta}^*_{lk}$. However, since $\tilde{\theta}^*\in\Theta^*_I$, we have that $\sum_l\tilde{\theta}^*_{lk}=P(M(1)=k)$, and hence (23) holds. Likewise, we have that $P^*(M(0)=k)=\sum_l\tilde{\theta}^*_{kl}=P(M(0)=k)$, so (24) holds. Observe further that if $f^{**}_{Y(1,k)|G=lk}=f^*_{Y(1,k)|G=lk}$ and $f^{**}_{Y(0,k)|G=kl}=f^*_{Y(0,k)|G=kl}$ for all l,k such that $\tilde{\theta}^*_{lk}>0$, then (25) and (26) hold by assumption. To complete the proof of the lemma, it thus suffices to show that we can construct joint densities $f^*_{Y(1,m),Y(0,m)|G}$ satisfying the conditions of Proposition 5.1 part 2 with marginals matching $f^*_{Y(1,k)|G=lk}$ and $f^*_{Y(0,k)|G=kl}$ for all l,k such that $\tilde{\theta}^*_{lk}>0$.

Note that for $l\neq k$, $\mathcal{F}_0$ and $\mathcal{F}_1$ depend on $(f^*_{Y(1,m),Y(0,m)|G=lk})_{m=1}^K$ only through $f^*_{Y(1,k)|G=lk}$ and $f^*_{Y(0,l)|G=lk}$. It thus suffices to choose

$$f^*_{Y(1,m),Y(0,m)|G=lk}(y_1,y_0)\propto f^*_{Y(1,k)|G=lk}(y_1)\cdot 1\{y_1=y_0\}\text{ if }m\neq l$$
$$f^*_{Y(1,m),Y(0,m)|G=lk}(y_1,y_0)\propto f^*_{Y(0,l)|G=lk}(y_0)\cdot 1\{y_1=y_0\}\text{ if }m=l,$$

which matches the required marginals and ensures that $P^*(Y(1,m)=Y(0,m)\,|\,G=lk)=1$ for all m.[32]

[32] The densities in the previous display are measurable with respect to the dominating measure $\tilde{\mu}$ such that for $B\subset\mathcal{Y}^2$, $\tilde{\mu}(B)=\mu(\{y:(y,y)\in B\})$.

Next, consider the case where $G = kk$. Observe that $\mathcal{F}_0$ and $\mathcal{F}_1$ do not depend at all on $f^*_{Y(1,m),Y(0,m)|G=kk}$ for $m \neq k$, and hence for $m \neq k$ we can choose $f^*_{Y(1,m),Y(0,m)|G=kk}$ corresponding to any arbitrary density (with marginals measurable wrt μ) such that $P^*(Y(1,m) = Y(0,m) \mid G = kk) = 1$. Now, suppose first that $\tilde{\theta}^*_{kk} > 0$. Recall that for scalar random variables H_1 and H_2 with densities h_1 and h_2 (measurable wrt μ), respectively, there exists a coupling $(H_1,H_2) \sim Q$ with marginals matching H_1 and H_2 such that $P_Q(H_1 \neq H_2) = tv$, where $tv = \int_{\mathcal{Y}} (h_1 - h_2)_+ d\mu$ is the total variation distance between H_1 and H_2. However, by assumption, $\int (f^*_{Y(1,k)|G=kk} - f^*_{Y(0,k)|G=kk})_+ d\mu = \eta_k / \tilde{\theta}^*_{kk}$, and hence there exists a joint density $f^*_{Y(0,k),Y(1,k)|G=kk}$ with marginals $f^*_{Y(1,k)|G=kk}$ and $f^*_{Y(0,k)|G=kk}$ such that $P^*(Y(1,k) \neq Y(0,k) \mid G = kk) = \eta_k / \tilde{\theta}^*_{kk}$, and hence (16) holds. Finally, suppose that $\tilde{\theta}^*_{kk} = 0$. We claim that in this case $\eta_k = 0$. Observe that since $\tilde{\theta}^* \in \Theta^*_I$ and $\tilde{\theta}^*_{kk} = 0$, we have that $\sum_{l:l \neq k} \tilde{\theta}^*_{lk} = P(M(1) = k)$. However,

$$\sup_A \left\{ P(Y^{\text{tot}}(1) \in A, M(1) = k) - P(Y^{\text{tot}}(0) \in A, M(0) = k) \right\} \leqslant \sup_A \left\{ P(Y^{\text{tot}}(1) \in A, M(1) = k) \right\} = P(M(1) = k),$$

and hence

$$\sup_A \left\{ P(Y^{\text{tot}}(1) \in A, M(1) = k) - P(Y^{\text{tot}}(0) \in A, M(0) = k) \right\} - \sum_{l:l \neq k} \tilde{\theta}^*_{lk} \leqslant 0,$$

which implies that $\eta_k = 0$. It follows that (16) holds (with both sides of the equation equal to zero) regardless of the value of $P^*(Y(1,m_k) \neq Y(0,m_k) \mid G = kk)$. It thus suffices to choose $f^*_{Y(0,k),Y(1,k)|G=kk}$ to be any joint distribution with marginals $f^*_{Y(0,k)|G=kk}$ and $f^*_{Y(1,k)|G=kk}$ (e.g. using the perfect-dependence copula). ▲

We now show that there exist densities satisfying the conditions of the Lemma. (For the remainder of the proof, all densities are measurable wrt μ, and integrals are taken wrt μ, so we omit the dependence for ease of notation.) Fix k. First, suppose that $\tilde{\theta}^*_{kk} = 0$. Choose

$$f^*_{Y(1,k)|G=lk} = f_{Y^{\text{tot}}(1)|M(1)=k} \text{ for all } l$$
$$f^*_{Y(0,k)|G=kl} = f_{Y^{\text{tot}}(0)|M(0)=k} \text{ for all } l.$$

Note that

$$\sum_l \tilde{\theta}^*_{lk} f^*_{Y(1,k)|G=lk} = \left(\sum_l \tilde{\theta}^*_{lk} \right) f_{Y^{\text{tot}}(1)|M(1)=k} = P(M(1) = k) \cdot f_{Y^{\text{tot}}(1)|M(1)=k} = f_{Y^{\text{tot}}(1),M(1)=k},$$

where the first equality uses the construction of f^*, the second equality uses the fact that $\tilde{\theta}^* \in \Theta^*_I$, and the final equality uses the definition of the partial density. It follows that (20) holds. Equation (21) can be verified analogously. Since $\tilde{\theta}^*_{kk} = 0$, we do not need to verify (22).

Now, suppose that $\tilde{\theta}^*_{kk} > 0$. Assume first that $\eta_k > 0$. Define $f_{min} := \min\{f_{Y^{\text{tot}}(1),M(1)=k}, f_{Y^{\text{tot}}(0),M(0)=k}\}$. Suppose first that $f_{min} = 0$ (μ-a.e.). Consider the densities of the potential outcomes

$$f^*_{Y(1,k)|G=g} = f_{Y^{\text{tot}}(1),M(1)=k}/P(M(1)=k) \text{ for all } g$$
$$f^*_{Y(0,k)|G=g} = f_{Y^{\text{tot}}(0),M(0)=k}/P(M(0)=k) \text{ for all } g.$$

By construction, the densities are non-negative and integrate to 1, and thus are valid densities. Since the definition of Θ^*_I implies that $\sum_l \tilde{\theta}^*_{lk} = P(M(1)=k)$ and $\sum_l \tilde{\theta}^*_{kl} = P(M(0)=k)$, it is immediate that (20) and (21) hold. Moreover, since $f_{min} = 0$, it follows that $f_{Y^{\text{tot}}(0),M(0)=k} = 0$ whenever $f_{Y^{\text{tot}}(1),M(1)=k} > 0$, and consequently $(f^*_{Y(1,k)|G=kk} - f^*_{Y(0,k)|G=kk})_+ = f^*_{Y(1,k)|G=kk}$. It follows that

$$\tilde{\theta}^*_{kk} \int_{\mathcal{Y}} (f^*_{Y(1,k)|G=kk} - f^*_{Y(0,k)|G=kk})_+ = \tilde{\theta}^*_{kk} \int_{\mathcal{Y}} f^*_{Y(1,k)|G=kk} = \tilde{\theta}^*_{kk}.$$

Note, however, that

$$\begin{aligned}
\eta_k =& \int_{\mathcal{Y}} \left(f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k}\right)_+ - \sum_{l:l\neq k} \tilde{\theta}^*_{lk} \\
=& \int_{\mathcal{Y}} f_{Y^{\text{tot}}(1),M(1)=k} - \sum_{l:l\neq k} \tilde{\theta}^*_{lk} \\
=& P(M(1)=k) - \sum_{l:l\neq k} \tilde{\theta}^*_{lk} \\
=& \tilde{\theta}^*_{kk}
\end{aligned}$$

where the first equality uses the fact that $\sup_A \Delta^*_k(A) = \int_{\mathcal{Y}} \left(f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k}\right)_+$ and the assumption that $\eta_k > 0$; the second equality uses the fact that $f_{min} = 0$, and thus $f_{Y^{\text{tot}}(0),M(0)=k}$ is zero whenever $f_{Y^{\text{tot}}(1),M(1)=k} > 0$ as argued above; and the final equality uses the fact that $P(M(1)=k) = \tilde{\theta}^*_{kk} + \sum_{l:l\neq k} \tilde{\theta}^*_{lk}$ by the definition of the identified set Θ^*_I. It follows from the previous two displays that (22) holds.

Next, suppose that $\eta_k > 0$ and that $f_{min} > 0$ on a set of positive measure (wrt μ). Then $\int_{\mathcal{Y}} f_{min} > 0$, and since $f_{min} \geqslant 0$ by construction, it follows that $\tilde{f}_{min} = f_{min}/\int_{\mathcal{Y}} f_{min}$ is a valid density. Define $f_d := f_{Y^{\text{tot}}(d),M(d)=k} - f_{min}$ and $\tilde{f}_d := f_d/\int_{\mathcal{Y}} f_d$. We claim that the $\tilde{f}_d$ are valid densities. First, observe from the definition of f_{min} that $f_d \geqslant 0$ everywhere. To show that $\tilde{f}_d$ is a valid density, it thus remains to show that $\int_{\mathcal{Y}} f_d > 0$, in which case $\tilde{f}_d$ is well-defined and integrates to 1 by construction. Observe, however, that since $\eta_k > 0$,

$$0 < \sup_A \Delta^*_k(A) = \int_{\mathcal{Y}} (f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k})_+ = \int_{\mathcal{Y}} f_1$$

where the second equality follows from the fact that $(A-B)_+ = A - \min\{A,B\}$ and the definition of f_1. We thus see that $\int_{\mathcal{Y}} f_1 > 0$. Next, observe that if $f_0 = 0$ (μ-a.e.), then $(f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k})_+ = f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k}$ (μ-a.e.), and thus

$$\begin{aligned}\sup_A \Delta_k^*(A) &= \int_{\mathcal{Y}} f_{Y^{\text{tot}}(1),M(1)=k} - f_{Y^{\text{tot}}(0),M(0)=k} \\ &= P(M(1)=k) - P(M(0)=k) \\ &= \sum_{l:l\neq k} \tilde{\theta}_{lk}^* - \sum_{l:l\neq k} \tilde{\theta}_{kl}^* \\ &\leqslant \sum_{l:l\neq k} \tilde{\theta}_{lk}^*,\end{aligned}$$

which implies that $\sup_A \Delta_k^*(A) - \sum_{l:l\neq k} \tilde{\theta}_{lk}^* \leqslant 0$, which contradicts the assumption that $\eta_k > 0$. Hence, we see that $f_0 > 0$ on a set of positive measure (wrt μ), and thus $\int_{\mathcal{Y}} f_0 > 0$, completing the proof that the $\tilde{f}_d$ are valid densities. Now, let $\nu_k^* = \eta_k / \tilde{\theta}_{kk}^*$, and construct the densities as follows:

$$\begin{aligned}&f_{Y(d,k)|G=kk}^* = (1-\nu_k^*)\tilde{f}_{min} + \nu_k^* \tilde{f}_d \text{ for } d=0,1 \\ &f_{Y(1,k)|G=g}^* = \tilde{f}_1 \text{ for } g \in \{lk : l \neq k\} \\ &f_{Y(0,k)|G=g}^* = \tilde{f}_0 \text{ for } g \in \{kl : l \neq k\}.\end{aligned}$$

To verify that $f_{Y(d,k)|G=kk}^*$ is a valid density, we will show that $\nu_k^* \in [0,1]$, in which case $f_{Y(d,k)|G=kk}$ is a convex combination of valid densities and hence a valid density. Note that $\nu_k^* = \eta_k / \tilde{\theta}_{kk}^*$, where $\eta_k > 0$ and $\tilde{\theta}_{kk}^* > 0$ by assumption, from which we see that $\nu_k^* \geqslant 0$. To show that $\nu_k^* \leqslant 1$, observe that

$$\begin{aligned}\nu_k^* &= \frac{\sup_A \{P(Y^{\text{tot}}(1) \in A, M(1)=k) - P(Y^{\text{tot}}(0) \in A, M(0)=k)\} - \sum_{l:l\neq k} \tilde{\theta}_{lk}^*}{\tilde{\theta}_{kk}^*} \\ &\leqslant \frac{\sup_A \{P(Y^{\text{tot}}(1) \in A, M(1)=k)\} - \sum_{l:l\neq k} \tilde{\theta}_{lk}^*}{\tilde{\theta}_{kk}^*} \\ &= \frac{P(M(1)=k) - \sum_{l:l\neq k} \tilde{\theta}_{lk}^*}{\tilde{\theta}_{kk}^*} \\ &= \frac{\tilde{\theta}_{kk}^*}{\tilde{\theta}_{kk}^*}.\end{aligned}$$

We have thus verified that the density for $f_{Y(d,k)|G=kk}$ is valid.

We now verify that the specified densities satisfy (20). Note that

$$\sum_l \tilde{\theta}_{lk}^* f_{Y(1,k)|G=lk} = \left(\sum_{l:l\neq k} \tilde{\theta}_{lk}^* + \tilde{\theta}_{kk}^* \nu_k^* \right) \frac{f_1}{\int_{\mathcal{Y}} f_1} + \tilde{\theta}_{kk}^* (1-\nu_k^*) \frac{f_{min}}{\int_{\mathcal{Y}} f_{min}}.$$

Since $f_1+f_{min}=f_{Y^{\text{tot}}(1),M(1)=k}$ by the definition of f_1, to verify (20) it suffices to verify that $\left(\sum_{l:l\neq k}\tilde{\theta}^*_{lk}+\tilde{\theta}^*_{kk}\nu^*_k\right)/\int_{\mathcal{Y}}f_1=1$ and $\tilde{\theta}^*_{kk}(1-\nu^*_k)/\int_{\mathcal{Y}}f_{min}=1$. Observe, however, that

$$\begin{aligned}
\nu^*_k &= \frac{1}{\tilde{\theta}^*_{kk}}\left(\sup_A \Delta^*_k(A)-\sum_{l:l\neq k}\tilde{\theta}^*_{lk}\right)\\
&= \frac{1}{\tilde{\theta}^*_{kk}}\left(\int_{\mathcal{Y}}(f_{Y^{\text{tot}}(1),M(1)=k}-f_{min})-\sum_{l:l\neq k}\tilde{\theta}^*_{lk}\right)\\
&= \frac{1}{\tilde{\theta}^*_{kk}}\left(P(M(1)=k)-\int_{\mathcal{Y}}f_{min}-\sum_{l:l\neq k}\tilde{\theta}^*_{lk}\right)\\
&= \frac{1}{\tilde{\theta}^*_{kk}}\left(\tilde{\theta}^*_{kk}-\int_{\mathcal{Y}}f_{min}\right)=1-\frac{\int_{\mathcal{Y}}f_{min}}{\tilde{\theta}^*_{kk}}
\end{aligned}$$

where the first equality uses the definition of ν^*_k and the assumption that $\eta_k>0$; the second equality uses the fact that $\int_{\mathcal{Y}}(f-g)_+=\int_{\mathcal{Y}}(f-\min\{f,g\})$; the third equality uses basic properties of densities; and the fourth equality uses the fact that $P(M(1)=k)=\sum_{l:l\neq k}\tilde{\theta}^*_{lk}+\tilde{\theta}^*_{kk}$ since $\tilde{\theta}^*\in\Theta^*_I$. It is then immediate from the previous display that $\tilde{\theta}^*_{kk}(1-\nu^*_k)/\int_{\mathcal{Y}}f_{min}=1$. To show that $\left(\sum_{l:l\neq k}\tilde{\theta}^*_{lk}+\tilde{\theta}^*_{kk}\nu^*_k\right)/\int_{\mathcal{Y}}f_1=1$, we again use the fact that $P(M(1)=k)=\sum_{l:l\neq k}\tilde{\theta}^*_{lk}+\tilde{\theta}^*_{kk}$, to obtain that $\sum_{l:l\neq k}\tilde{\theta}^*_{lk}+\tilde{\theta}^*_{kk}\nu^*_k=P(M(1)=k)-(1-\nu^*_k)\tilde{\theta}^*_{kk}=P(M(1)=k)-\int_{\mathcal{Y}}f_{min}$, where the second equality uses the result in the previous display. However, from the definition of f_1, $\int_{\mathcal{Y}}f_1=\int_{\mathcal{Y}}f_{Y^{\text{tot}}(1),M(1)=k}-f_{min}=P(M(1)=k)-\int_{\mathcal{Y}}f_{min}$, and we thus see that $\left(\sum_{l:l\neq k}\tilde{\theta}^*_{lk}+\tilde{\theta}^*_{kk}\nu^*_k\right)/\int_{\mathcal{Y}}f_1=1$, as needed to verify (20). An analogous argument can be used to verify (21).

To show that the specified densities match (22), note that the construction of

$$\tilde{f}_d\propto f_{Y^{\text{tot}}(d),M(d)=k}-f_{min}$$

implies that $\tilde{f}_0=0$ whenever $(\tilde{f}_1-\tilde{f}_0)_+>0$. It follows that $\int_{\mathcal{Y}}(\tilde{f}_1-\tilde{f}_0)_+=\int_{\mathcal{Y}}\tilde{f}_1=1$. Hence,

$$\tilde{\theta}^*_{kk}\int_{\mathcal{Y}}(f^*_{Y(1,k)|G=kk}-f^*_{Y(0,k)|G=kk})_+=\tilde{\theta}^*_{kk}\int_{\mathcal{Y}}\nu^*_k(\tilde{f}_1-\tilde{f}_0)_+=\tilde{\theta}^*_{kk}\nu^*_k=\eta_k,$$

as needed.

Next, consider the case where $\eta_k=0$ and hence $\sup_A\Delta^*_k(A)-\sum_{l:l\neq k}\tilde{\theta}^*_{lk}\leqslant 0$. Consider the densities

$$f^*_{Y(1,k)|G=kk} = f^*_{Y(0,k)|G=kk} = f_{min} / \int_{\mathcal{Y}} f_{min}$$

$$f^*_{Y(1,k)|G=g} = \frac{1}{\sum_{l:l \neq k} \tilde{\theta}^*_{lk}} \left(f_{Y^{\text{tot}}(1),M(1)=k} - \tilde{\theta}^*_{kk} \frac{f_{min}}{\int_{\mathcal{Y}} f_{min}} \right) \text{ for all } g \in \{lk : l \neq k\}$$

$$f^*_{Y(0,k)|G=g} = \frac{1}{\sum_{l:l \neq k} \tilde{\theta}^*_{kl}} \left(f_{Y^{\text{tot}}(0),M(0)=k} - \tilde{\theta}^*_{kk} \frac{f_{min}}{\int_{\mathcal{Y}} f_{min}} \right) \text{ for all } g \in \{kl : l \neq k\}.$$

We now verify that the specified densities are in fact proper. First, we showed above that if $f_{min} = 0$ (μ-a.e.), then $\eta_k = \tilde{\theta}^*_{kk} > 0$. Hence, since $\eta_k = 0$, it must be that $\int_{\mathcal{Y}} f_{min} > 0$, so that $f_{min} / \int_{\mathcal{Y}} f_{min}$ is a proper density. Next, we verify that the specified densities for $g \neq kk$ are non-negative. Recall that by assumption $\sup_A \Delta^*_k(A) - \sum_{l:l \neq k} \tilde{\theta}^*_{lk} \leqslant 0$. Note, further, that

$$\sup_A \Delta^*_k(A) = \int_{\mathcal{Y}} f_{Y^{\text{tot}}(1),M(1)=k} - f_{min} = P(M(1)=k) - \int_{\mathcal{Y}} f_{min},$$

and hence

$$P(M(1)=k) - \int_{\mathcal{Y}} f_{min} - \sum_{l:l \neq k} \tilde{\theta}^*_{lk} \leqslant 0.$$

However, since $P(M(1)=k) - \sum_{l:l \neq k} \tilde{\theta}^*_{lk} = \tilde{\theta}^*_{kk}$ by the definition of Θ^*_I, we see from the previous display that $\int_{\mathcal{Y}} f_{min} \geqslant \tilde{\theta}^*_{kk}$, and thus $\frac{\tilde{\theta}^*_{kk}}{\int_{\mathcal{Y}} f_{min}} \leqslant 1$. But since $f_{Y^{\text{tot}}(d),M(d)=k} \geqslant f_{min}$ by construction, it follows that $f_{Y^{\text{tot}}(d),M(d)=k} - \frac{\tilde{\theta}^*_{kk}}{\int_{\mathcal{Y}} f_{min}} f_{min} \geqslant 0$, and hence the specified densities for $f_{Y(d,k)|G=g}$ for $g \neq kk$ are non-negative. To see that these densities integrate to 1, observe that

$$\int_{\mathcal{Y}} \left(f_{Y^{\text{tot}}(1),M(1)=k} - \frac{\tilde{\theta}^*_{kk}}{\int_{\mathcal{Y}} f_{min}} f_{min} \right) = P(M(1)=k) - \tilde{\theta}^*_{kk} = \sum_{l:l \neq k} \tilde{\theta}^*_{lk}$$

and similarly

$$\int_{\mathcal{Y}} \left(f_{Y^{\text{tot}}(0),M(0)=k} - \frac{\tilde{\theta}^*_{kk}}{\int_{\mathcal{Y}} f_{min}} f_{min} \right) = P(M(0)=k) - \tilde{\theta}^*_{kk} = \sum_{l:l \neq k} \tilde{\theta}^*_{kl}.$$

Finally, it is trivial to verify from the construction of the densities above that equations (20), (21), and (22) hold. ▲

Proof of Proposition 3.1 Under Assumption 1, $(Y,M) \mid D = d \sim (Y^{\text{tot}}(d), M(d))$ for $d = 0,1$ (where recall $Y^{\text{tot}}(d) := Y(d, M(d))$). The result then follows immediately from Proposition 5.1. Specifically, as noted in the main text, under Assumption 1, $\Delta_k(A) = \Delta^*_k(A)$ and $\Theta_I = \Theta^*_I$.

Hence, (14) from Proposition 5.1 implies (4), and likewise (15) implies (5), which yields the first part of the Proposition. For the second part, for any $\tilde{\theta} \in \Theta_I = \Theta_I^*$, part 2 of Proposition 5.1 implies that there exists a distribution P^* for $(Y(\cdot,\cdot),M(\cdot))$ that is consistent with the marginals $(Y^{\text{tot}}(d),M(d)) \sim (Y,M) \mid D=d$ for $d=0,1$ such that

$$\tilde{\theta}_{kk} P^*(Y(1,k) \neq Y(0,k) \mid G=kk) = \left(\sup_A \Delta_k^*(A) - \sum_{l:l \neq k} \tilde{\theta}_{lk} \right)_+ = \left(\sup_A \Delta_k(A) - \sum_{l:l \neq k} \tilde{\theta}_{lk} \right)_+$$

and $P^*(Y(1,m) \neq Y(0,m) \mid G=lk) = 0$ if either $l \neq k$ or $m \notin \{l,k\}$. Part 2 of the Proposition is thus satisfied for $P^\dagger$ the distribution over $(Y(\cdot,\cdot),M(\cdot),D)$ defined such that $(Y(\cdot,\cdot),M(\cdot)) \sim P^*$, $D \sim P_D$ (where P_D is the marginal distribution of D under P_{obs}), and $D \perp\!\!\!\perp (Y(\cdot,\cdot),M(\cdot))$. ▲

Proof of Lemma 3.1 To prove Lemma 3.1, we prove the following result, which generalizes the bounds given in Lemma 3.1 to the case where Y may not be continuously distributed. For notation, for a distribution F, let $F^{-1}(u) = \inf\{y : F(y) \geqslant u\}$ be the uth quantile of F.

Lemma A.2. *Suppose Assumption 1 holds. Then if $\check{\theta}^1_{kk} > 0$,*

$$\frac{1}{\check{\theta}^1_{kk}} \int_0^{\check{\theta}^1_{kk}} F^{-1}_{Y|D=1,M=k}(u) du \leqslant E[Y(1,k) \mid G=kk] \leqslant \frac{1}{\check{\theta}^1_{kk}} \int_{1-\check{\theta}^1_{kk}}^1 F^{-1}_{Y|D=1,M=k}(u)$$

and if $\check{\theta}^0_{kk} > 0$,

$$\frac{1}{\check{\theta}^0_{kk}} \int_0^{\check{\theta}^0_{kk}} F^{-1}_{Y|D=0,M=k}(u) du \leqslant E[Y(0,k) \mid G=kk] \leqslant \frac{1}{\check{\theta}^0_{kk}} \int_{1-\check{\theta}^0_{kk}}^1 F^{-1}_{Y|D=0,M=k}(u).$$

The bounds are sharp in the sense that there exists a distribution $P^\dagger$ for $(Y(\cdot,\cdot),M(\cdot),D)$ consistent with Assumption 1 and the observable data and with $\theta_{lk} = P^\dagger(G=lk)$ such that the bounds hold with equality. If the distributions of $Y \mid D=d, M=k$ are continuous, then the bounds can equivalently be written as

$$E[Y \mid M=k, D=1, Y \leqslant y^1_{\check{\theta}^1_{kk}}] \leqslant E[Y(1,k) \mid G=kk] \leqslant E[Y \mid M=k, D=1, Y \geqslant y^1_{1-\check{\theta}^1_{kk}}]$$

and

$$E[Y \mid M=k, D=0, Y \leqslant y^0_{\check{\theta}^0_{kk}}] \leqslant E[Y(0,k) \mid G=kk] \leqslant E[Y \mid M=k, D=0, Y \geqslant y^0_{1-\check{\theta}^0_{kk}}],$$

where $y^d_q := F^{-1}_{Y|D=d,M=k}(q)$ is the qth quantile of $Y \mid D=d, M=k$.

Proof. We begin by deriving the bounds for $E[Y(1,k)\,|\,G=kk]$. Observe that under Assumption 1,

$$F_{Y|D=1,M=k}=\check{\theta}^1_{kk}F_{Y(1,k)|G=kk}+(1-\check{\theta}^1_{kk})H,$$

where $H=\frac{1}{\sum_{l:l\neq k}\theta_{lk}}\sum_{l:l\neq lk}\theta_{lk}F_{Y(1,k)|G=lk}$ is a valid CDF (corresponding to a mixture of the distributions of $Y(1,k)\,|\,G=g$ for types $g=lk$, $l\neq k$). Hence,

$$F_{Y(1,k)|G=kk}=\frac{1}{\check{\theta}^1_{kk}}F_{Y|D=1,M=k}-\frac{1-\check{\theta}^1_{kk}}{\check{\theta}^1_{kk}}H.$$

From the fact that CDFs are bounded between 0 and 1, it follows that

$$\max\left\{\frac{1}{\check{\theta}^1_{kk}}F_{Y|D=1,M=k}-\frac{1-\check{\theta}^1_{kk}}{\check{\theta}^1_{kk}},0\right\}\leqslant F_{Y(1,k)|G=kk}\leqslant\min\left\{\frac{1}{\check{\theta}^1_{kk}}F_{Y|D=1,M=k},1\right\}$$

Recall that if $F_1\leqslant F_2$ everywhere for CDFs F_1 and F_2, the F_1 distribution first-order stochastically dominates the F_2 distribution, and thus $E_{F_1}[Y]\geqslant E_{F_2}[Y]$. Hence, we have that $E_{F_{ub}}[Y(1,k)]\leqslant E[Y(1,k)\,|\,G=kk]\leqslant E_{F_{lb}}[Y(1,k)]$, where F_{lb},F_{ub} are respectively the lower and upper bounds on the CDF given in the previous display.

Now, let U be uniform on $[0,1]$, and consider the random variable $Y_{ub}\sim F^{-1}_{Y|D=1,M=k}(U)\,|\,U\in[0,\check{\theta}^1_{kk}]$. Observe that

$$\begin{aligned}F_{Y_{ub}}(y)&=P(F^{-1}_{Y|D=1,M=k}(U)\leqslant y\,|\,U\in[0,\check{\theta}^1_{kk}])\\&=P(F_{Y|D=1,M=k}(y)\geqslant U\,|\,U\in[0,\check{\theta}^1_{kk}])\\&=\min\left\{\frac{1}{\check{\theta}^1_{kk}}F_{Y|D=1,M=k}(y),1\right\}=F_{ub}(y).\end{aligned}$$

It follows that $E_{F_{ub}}[Y(1,k)]=E[F^{-1}_{Y|D=1,M=k}(U)\mid U\in[0,\check{\theta}^1_{kk}]]=\frac{1}{\check{\theta}^1_{kk}}\int_0^{\check{\theta}^1_{kk}}F^{-1}_{Y|D=1,M=k}(u)du$, which gives the lower-bound on $E[Y(1,k)\mid G=kk]$ given in the lemma. When Y is continuously distributed, note that $Y_{ub}\sim\left(Y\,|\,D=1,M=k,Y\leqslant y_{\check{\theta}^1_{kk}}\right)$, and thus we can also write the lower-bound as $E[Y\mid D=1,M=k,Y\leqslant y_{\check{\theta}^1_{kk}}]$. Analogously, we can verify that the random variable $Y_{lb}\sim F^{-1}_{Y|D=1,M=k}(U)\mid U\in[1-\check{\theta}^1_{kk},1]$ has the CDF F_{lb}, which gives the upper bound on $E[Y(1,k)\,|\,G=kk]$ given in the proposition.

To show that the lower bound is sharp, consider $P^{\dagger}$ such that $D\perp\!\!\!\perp Y(\cdot,\cdot),M(\cdot)$ and $P^{\dagger}(M(0)=l,M(1)=k)=\theta_{lk}$, and the marginal distributions of the potential outcomes are such that $Y(1,k)\,|\,G=kk\sim Y_{lb}$ and $Y(1,k)\,|\,G=lk\sim F^{-1}_{Y|D=1,M=k}(U)\,|\,U\in[\check{\theta}^1_{kk},1]$ for all $g=lk$ with

$l \neq k$. Then the distribution of $Y \mid M=k, D=1$ is given by the mixture:

$$\check{\theta}^1_{kk}\left(F^{-1}_{Y|D=1,M=k}(U) \mid U \in [0,\check{\theta}^1_{kk}]\right) + (1-\check{\theta}^1_{kk})\left(F^{-1}_{Y|D=1,M=k}(U) \mid U \in [\check{\theta}^1_{kk},1]\right) \sim F^{-1}_{Y|D=1,M=k}(U).$$

Recalling that if Y has CDF F, then $Y \sim F^{-1}(U)$, we see that the implied distribution of $Y \mid M=k, D=1$ under $P^\dagger$ matches the observable data. (The marginals of $Y(0,m) \mid G$ under $P^\dagger$ can be chosen to be any set of distributions matching the observable data; likewise the copula of potential outcomes can be chosen arbitrarily.) The sharpness of the upper bound can be shown analogously. Sharp bounds for $E[Y(0,k) \mid G=kk]$ can be shown analogously to those for $E[Y(1,k) \mid G=kk]$. □

Proof of Proposition 3.2

Proof. From Lemma A.2, we have that

$$\begin{aligned}&\inf_{\check{\theta}^d_{kk} \in \check{\Theta}_{I,d}} E[F^{-1}_{Y|D=d,M=k}(U) \mid U \in [0,\check{\theta}^d_{kk}]]\\ \leqslant & E[Y(1,k) \mid G=kk]\\ \leqslant & \sup_{\check{\theta}^d_{kk} \in \check{\Theta}_{I,d}} E[F^{-1}_{Y|D=d,M=k}(U) \mid U \in [1-\check{\theta}^d_{kk},1]]\end{aligned}$$

for U uniformly distributed and $\check{\Theta}_{I,d}$ the set of values for $\check{\theta}^d_{kk} = \tilde{\theta}_{kk}/P(M=k \mid D=d)$ consistent with $\tilde{\theta} \in \Theta_I$. Since $F^{-1}_{Y|D=d,M=k}(U)$ is increasing in U, it follows that the inf and sup are both obtained at $\check{\theta}^{min}_{kk}$. The bounds for $ADE_k = E[Y(1,k) - Y(0,k) \mid G=kk]$ follow simply from differencing the bounds for the two potential outcomes in the previous display. Sharpness for the bounds for ADE_k follows from the fact that, as shown in the proof to Lemma A.2, for each $d=0,1$, the bounds for $E[Y(d,k) \mid G=kk]$ can be achieved only by specifying the marginals of $Y(d,k) \mid G=kk$ (and choosing the remaining potential outcomes in any arbitrary way that matches the data) and thus the bounds for $d=0,1$ can be achieved simultaneously. □

B Additional Theoretical Results

B.1 Minimum value of ν_k achieved at θ^{min}_{kk}

The following result formalizes the sense in which the lower bound on ν_k implied by (5) from Proposition 3.1 is achieved at the minimum possible value of $\tilde{\theta}_{kk}$ in the identified set, $\tilde{\theta}^{min}_{kk} := \inf_{\tilde{\theta} \in \Theta_I} \tilde{\theta}_{kk}$.

Lemma B.1. *If $\tilde{\theta}_{kk}^{min} > 0$, then*

$$\inf_{\substack{\nu_k \geqslant 0, \tilde{\theta} \in \Theta_I \\ \textit{s.t. } (5) \textit{ holds}}} \nu_k = \frac{1}{\tilde{\theta}_{kk}^{min}} \left(\sup_A \Delta_k(A) - P(M = m_k \mid D = 1) + \tilde{\theta}_{kk}^{min} \right)_+ .$$

On the other hand, if $\tilde{\theta}_{kk}^{min} = 0$, then

$$\inf_{\substack{\nu_k \geqslant 0, \tilde{\theta} \in \Theta_I \\ \textit{s.t. } (5) \textit{ holds}}} \nu_k = 0.$$

Proof. For simplicity of notation, without loss of generality let $m_k = k$. We argued in the main text that (5) can be equivalently written as

$$\tilde{\theta}_{kk} \nu_k \geqslant \left(\sup_A \Delta_k(A) - (P(M = k \mid D = 1) - \tilde{\theta}_{kk}) \right)_+ .$$

Hence, when $\tilde{\theta}_{kk}^{min} > 0$, we have that

$$\begin{aligned} \inf_{\substack{\nu_k \geqslant 0, \tilde{\theta} \in \Theta_I \\ \text{s.t. } (5) \text{ holds}}} \nu_k &= \inf_{\tilde{\theta} \in \Theta_I} \frac{1}{\tilde{\theta}_{kk}} \left(\sup_A \Delta_k(A) - (P(M = k \mid D = 1) - \tilde{\theta}_{kk}) \right)_+ \\ &= \inf_{\tilde{\theta} \in \Theta_I} \left(\frac{1}{\tilde{\theta}_{kk}} \left(\sup_A \Delta_k(A) - P(M = k \mid D = 1) \right) + 1 \right)_+ . \end{aligned}$$

To show that the inf is achieved at $\tilde{\theta}_{kk}^{min}$, it thus suffices to show that

$$\sup_A \Delta_k(A) - P(M = k \mid D = 1) \leqslant 0,$$

in which case the expression inside the inf is weakly increasing in $\tilde{\theta}_{kk}$. Note, however, that

$$\begin{aligned} \sup_A \Delta_k(A) &= \sup_A \{ P(Y \in A, M = k \mid D = 1) - P(Y \in A, M = k \mid D = 0) \} \\ &\leqslant \sup_A P(Y \in A, M = k \mid D = 1) \\ &= P(M = k \mid D = 1), \end{aligned}$$

which completes the proof for the case where $\theta_{min}^{kk} > 0$.

Next, suppose that $\tilde{\theta}_{min}^{kk} = 0$. It is straightforward to verify that since R is closed by Assumption 2, Θ_I is also closed. It follows that there exists $\tilde{\theta}^* \in \Theta_I$ such that $\tilde{\theta}_{kk}^* = 0$. Since $\tilde{\theta}^* \in \Theta_I$, we have that $\sum_l \tilde{\theta}_{lk}^* = P(M = k \mid D = 1)$. Combined with the fact that $\tilde{\theta}_{kk}^* = 0$, we thus have that

$\sum_{l:l\neq k}\tilde{\theta}^*_{lk}=P(M=k\,|\,D=1)$. This combined with the inequality in the previous display implies that $(\sup_A \Delta_k(A)-\sum_{l:l\neq k}\tilde{\theta}^*_{lk})_+=0$, and hence (5) holds with $\tilde{\theta}=\tilde{\theta}^*$ and $\nu_k=0$. □

B.2 Closed-form solution for θ_{kk} with fully-ordered M

The following result formalizes the closed-form solution for $\tilde{\theta}^{min}_{kk}$ when M is fully-ordered and we impose monotonicity, as discussed in Remark 1.

Lemma B.2. *Suppose M is fully-ordered, so that $m_0<m_1<...<m_{K-1}$. Suppose Assumptions 1 and 2 are satisfied, where $R=\{\theta\in\Delta:\theta_{lk}=0 \text{ if } m_l>m_k\}$ imposes the monotonicity assumption that $M(1)\geqslant M(0)$. Then*

$$\tilde{\theta}_{kk}\geqslant P(M=m_k\,|\,D=1)-\min\{P(M=m_k\,|\,D=1),P(M\geqslant m_k\,|\,D=1)-P(M\geqslant m_k\,|\,D=0)\}$$

for all $\tilde{\theta}\in\Theta_I$, and there exists $\tilde{\theta}\in\Theta_I$ such that inequality holds with equality simultaneously for all k.

Proof. The result is an immediate corollary of Lemma B.3 below, since $\Theta_I=\Theta_I^*$ under Assumption 1. □

Lemma B.3. *Suppose M is fully-ordered, so that $m_0<m_1<...<m_{K-1}$. Suppose Assumption 2 holds with $R=\{\theta\in\Delta:\theta_{lk}=0 \text{ if } m_l>m_k\}$, which imposes the monotonicity assumption that $M(1)\geqslant M(0)$ (almost surely). Let Θ_I^* be as defined in (13). Then*

$$\tilde{\theta}^*_{kk}\geqslant P(M(1)=m_k)-\min\{P(M(1)=m_k),P(M(1)\geqslant m_k)-P(M(0)\geqslant m_k)\} \tag{27}$$

for all $\tilde{\theta}^*\in\Theta_I^*$, and there exists $\tilde{\theta}^*\in\Theta_I$ such that the inequality holds with equality simultaneously for all k.

Proof. For simplicity of notation, without loss of generality let $m_k=k$. Observe that the definition of Θ_I^* together with the assumed form for R implies that for any $\tilde{\theta}^*\in\Theta_I^*$,

$$P(M(1)=k)=\tilde{\theta}^*_{kk}+\sum_{l:l<k}\tilde{\theta}^*_{lk} \tag{28}$$

and hence the conclusion of the proposition holds if and only if

$$\sum_{l:l<k}\tilde{\theta}^*_{lk}\leqslant\min\{P(M(1)=k),P(M(1)\geqslant k)-P(M(0)\geqslant k)\}, \tag{29}$$

for $k=0,...,K-1$, and there exists some $\tilde{\theta}^*\in\Theta_I^*$ such that the inequality holds with equality for all k.

We first show the inequality in (29). It is immediate from (28) that

$$\sum_{l:l<k} \tilde{\theta}^*_{lk} \leqslant P(M(1)=k).$$

Moreover, using the restriction that $\tilde{\theta}^*_{lk}=0$ if $l>k$, we have that

$$P(M(1)\geqslant k)-P(M(0)\geqslant k)=\sum_{l:l<k}\sum_{k':k'\geqslant k}\tilde{\theta}^*_{lk'}\geqslant \sum_{l:l<k}\tilde{\theta}^*_{lk},$$

which together with the previous display gives the inequality in (29).

We next show there exists a $\tilde{\theta}^*\in\Theta^*_I$ that satisfies all of the inequalities with equality. To obey monotonicity, we set $\tilde{\theta}^*_{lk}=0$ whenever $k<l$.

We now recursively set the remaining $\tilde{\theta}^*_{lk}$. Start with $k=0$. Set $\tilde{\theta}^*_{00}=P(M(1)=0)$. Note that the monotonicity assumption that $M(1)\geqslant M(0)$ almost surely implies that $P(M(1)=0)\leqslant P(M(0)=0)$. It is then straightforward to verify that the following properties hold for $\bar{k}=0$ (in what follows, we interpret sums over empty sets as zero):

(i) $\sum_{l:l<j}\tilde{\theta}^*_{lj}=\min\{P(M(1)=j),P(M(1)\geqslant j)-P(M(0)\geqslant j)\}$ for all $j\leqslant\bar{k}$

(ii) $\sum_{l:l\leqslant j}\tilde{\theta}^*_{lj}=P(M(1)=j)$ for all $j\leqslant\bar{k}$

(iii) $\sum_{l:l\leqslant\bar{k}}\tilde{\theta}^*_{jl}\leqslant P(M(0)=j)$ for all $j\leqslant\bar{k}$.

Now, suppose that for some $k\geqslant 1$, $\tilde{\theta}^*_{lj}$ has been determined for all $l=0,\ldots,K-1$ and all $j=0,\ldots,k-1$, and properties (i)-(iii) hold for $\bar{k}=k-1$. (We showed above that this holds in the base case $k=1$.) Set $\tilde{\theta}^*_{kk}=P(M(1)=k)-\min\{P(M(1)=k),P(M(1)\geqslant k)-P(M(0)\geqslant k)\}$. For $l=0,\ldots,k-1$, proceed as follows

1. If $\sum_{l':l'<l}\tilde{\theta}^*_{l'k}=P(M(1)\geqslant k)-P(M(0)\geqslant k)$, then set $\tilde{\theta}^*_{lk}=0$.
2. Otherwise, set

$$\tilde{\theta}^*_{lk}=\min\left\{P(M(1)\geqslant k)-P(M(0)\geqslant k)-\sum_{l':l'<l}\tilde{\theta}^*_{l'k}\,,\,P(M(0)=l)-\sum_{k':k'<k}\tilde{\theta}^*_{lk'}\right\}.$$

Note that the first term in the minimum is weakly positive by construction while property (iii) ensures that the second term in the minimum is non-negative, so that $\tilde{\theta}^*_{lk}\geqslant 0$. We claim that the construction above implies that

$$\sum_{l:l<k}\tilde{\theta}^*_{lk}=\min\{P(M(1)=k),P(M(1)\geqslant k)-P(M(0)\geqslant k)\}.$$

To see why this is the case, suppose towards contradiction that

$$\sum_{l:l<k}\tilde{\theta}^*_{lk} < \min\{P(M(1)=k), P(M(1)\geqslant k)-P(M(0)\geqslant k)\}.$$

Then $\tilde{\theta}^*_{lk}$ is always set via step 2 in the procedure above. However, the construction of $\tilde{\theta}^*_{lk}$ in step 2 combined with the fact that $\sum_{l:l<k}\tilde{\theta}^*_{lk} < P(M(1)\geqslant k)-P(M(0)\geqslant k)$ implies that for all $l=0,...,k-1$, we have that

$$\tilde{\theta}^*_{lk} = P(M(0)=l)-\sum_{j:j<k}\tilde{\theta}^*_{lj}.$$

Summing over $l<k$, we obtain that

$$\begin{aligned}
\sum_{l:l<k}\tilde{\theta}^*_{lk} &= \sum_{l:l<k}P(M(0)=l)-\sum_{l:l<k}\sum_{j:j<k}\tilde{\theta}^*_{lj} \\
&= \sum_{l:l<k}P(M(0)=l)-\sum_{j:j<k}\sum_{l:l<k}\tilde{\theta}^*_{lj} && \text{(Reversing order of sums)} \\
&= \sum_{l:l<k}P(M(0)=l)-\sum_{j:j<k}\sum_{l:l\leqslant j}\tilde{\theta}^*_{lj} && \text{(Using monotonicity)} \\
&= \sum_{l:l<k}P(M(0)=l)-\sum_{j:j<k}P(M(1)=j) && \text{(Using property (ii))} \\
&= P(M(0)<k)-P(M(1)<k) \\
&= P(M(1)\geqslant k)-P(M(0)\geqslant k)
\end{aligned}$$

which is a contradiction.
It follows that property (i) holds also for $\bar{k}=k$. Likewise, the construction of $\tilde{\theta}^*_{kk}$ combined with property (i) implies that property (ii) holds for $\bar{k}=k$. Finally, the construction of $\tilde{\theta}^*_{lk}$ (particularly step 2) guarantees that property (iii) holds for $\bar{k}=k$ as well.

By induction we can obtain $\tilde{\theta}^*$ satisfying properties (i) through (iii) for $\bar{k}=K-1$. The resulting $\tilde{\theta}^*$ satisfies monotonicity and is bounded between 0 and 1 by construction. Property (ii) guarantees that $\tilde{\theta}^*$ matches the marginal of $M(1)$, i.e. $\sum_l\tilde{\theta}^*_{lk}=P(M(1)=k)$.

It thus remains only to establish that $\tilde{\theta}^*$ matches the marginal distribution of $M(0)$. Property (ii) implies that $\sum_l\tilde{\theta}^*_{jl}\leqslant P(M(0)=j)$. To establish equality for all j, it thus suffices to show that $\sum_j\sum_l\tilde{\theta}^*_{jl}\geqslant\sum_jP(M(0)=j)=1$. Note, however, that from property (ii) and monotonicity, we have

$$\sum_j\sum_l\tilde{\theta}^*_{jl}=\sum_j\left(\sum_{l:l\leqslant j}\tilde{\theta}^*_{lj}\right)=\sum_jP(M(1)=j)=1,$$

which completes the proof. □

B.3 Additional results for IV

Consider the setting in Section 5 in which we have a binary instrument Z for D that satisfies the Imbens and Angrist (1994) assumptions. Corollary 5.1 provides sharp testable implications of the sharp null that D affects Y only through M for instrument-compliers. Note, however, that the instrument exclusion restriction implies that Z affects Y only through D, and the sharp null implies that D affects Y only through M. Hence, under the sharp null, it follows that Z affects Y only through M. A simple alternative approach to testing the sharp null would therefore be to apply the testable implications derived under random assignment in Corollary 3.1 viewing Z as the randomized treatment and ignoring D. In this section, we show that the two approaches coincide when one imposes that $M(d)$ is monotonic in d (i.e. R is as given in (3)). Thus, if one is willing to impose monotonicity, one can simply apply our approach for experiments using Z as the treatment variable. If one does not impose monotonicity of $M(d)$, then this approach remains valid but potentially loses information relative to using Corollary 3.1.

To see why this is the case, recall from Section 5 that we can identify the marginal distributions $(Y^{\text{tot}}(d),M(d))\,|\,C^z=1$, where C^z is an indicator for instrument-compliers. We let Θ^*_{I,C^z} denote the analog to Θ^*_I defined in (13), with all probabilities conditional on $C^z=1$. Observe that Corollary 5.1 together with the formulas derived in Section 5 for $(Y^{\text{tot}}(d),M(d))\,|\,C^z=1$ imply that under the sharp null, there exists some $\tilde{\theta}^*\in\Theta^*_{I,C^z}$ such that for all $k=0,\dots,K-1$,

$$\sup_A \frac{\Delta^Z_k(A)}{\alpha_C} \leqslant \frac{E[D\cdot 1\{M=m_k\}\,|\,Z=1]-E[D\cdot 1\{M=m_k\}\,|\,Z=0]}{\alpha_C} - \tilde{\theta}^*_{kk} \tag{30}$$

where

$$\Delta^Z_k(A) := P(Y\in A, M=m_k\,|\,Z=1)-P(Y\in A, M=m_k\,|\,Z=0)$$

and

$$\alpha_C := E[D\,|\,Z=1]-E[D\,|\,Z=0]$$

is the identified share of instrument-compliers. Note that if M is fully-ordered and R imposes monotonicity, then by Lemma B.3,

$$\inf_{\tilde{\theta}\in\Theta^*_{I,C^z}} \tilde{\theta}^*_{kk} = (P(M(1)=m_k\,|\,C^z=1)-(P(M(1)\geqslant m_k\,|\,C^z=1)-P(M(0)\geqslant m_k\,|\,C^z=1))_+ =: \tilde{\theta}^{min,*}_{kk}$$

and there exists $\tilde{\theta}^*\in\Theta^*_{I,C^z}$ such that $\tilde{\theta}^*_{kk}=\tilde{\theta}^{min,*}_{kk}$ for all k. Using this result along with the expressions derived in Section 5 for $(Y^{\text{tot}}(d),M(d))\,|\,C^z=1$, it follows that under monotonicity,

(30) is equivalent to

$$\sup_A \frac{\Delta_k^Z(A)}{\alpha_C} \leqslant \frac{E[D\cdot 1\{M=m_k\}\,|\,Z=1]-E[D\cdot 1\{M=m_k\}\,|\,Z=0]}{\alpha_C}$$
$$-\left(\frac{E[D\cdot 1\{M=m_k\}\,|\,Z=1]-E[D\cdot 1\{M=m_k\}\,|\,Z=0]}{\alpha_C}-\frac{P(M\geqslant m_k\,|\,Z=1)-P(M\geqslant m_k\,|\,Z=0)}{\alpha_C}\right)_+$$

Multiplying through by α_C and using the fact that $a-(a-b)_+=\min\{a,b\}$ for any a,b, we obtain the equivalent implication

$$\sup_A \Delta_k^Z(A) \leqslant \min\{E[D\cdot 1\{M=m_k\}\,|\,Z=1]-E[D\cdot 1\{M=m_k\}\,|\,Z=0],$$
$$P(M\geqslant m_k\,|\,Z=1)-P(M\geqslant m_k\,|\,Z=0)\}. \tag{31}$$

On the other hand, we could ignore D and simply use Corollary 3.1 to test whether Z affects Y only through M (relabeling treatment 'D' with treatment 'Z' in Corollary 3.1). The testable implication is that there exists some $\tilde{\theta}\in\Theta_I$ such that for all $k=0,\ldots,K-1$

$$\sup_A \Delta_k^Z(A) \leqslant P(M=m_k\,|\,Z=1)-\tilde{\theta}_{kk}. \tag{32}$$

Using the expression for $\tilde{\theta}_{kk}^{min}$ under monotonicity given in Lemma B.2, this is equivalent to, for all $k=0,\ldots,K-1$,

$$\sup_A \Delta_k^Z(A) \leqslant P(M=m_k\,|\,Z=1)-(P(M=m_k\,|\,Z=1)-(P(M\geqslant m_k\,|\,Z=1)-P(M\geqslant m_k\,|\,Z=0)))_+$$
$$=\min\{P(M=m_k\,|\,Z=1),P(M\geqslant m_k\,|\,Z=1)-P(M\geqslant m_k\,|\,Z=0)\}. \tag{33}$$

We now claim that (31) and (33) are equivalent under our assumption that Z satisfies the Imbens and Angrist (1994) assumptions. Note that both inequalities take the form

$$\sup_A \Delta_k^Z(A) \leqslant \min\{q,P(M\geqslant m_k\,|\,Z=1)-P(M\geqslant m_k\,|\,Z=0)\} \tag{34}$$

but differ in the choice of q. Note that the upper bound is weakly increasing in q. Moreover, since $D\in\{0,1\}$,

$$E[D\cdot 1\{M=m_k\}\,|\,Z=1]\leqslant E[1\{M=m_k\}\,|\,Z=1]=P(M=m_k\,|\,Z=1).$$

We thus see that the upper bound in (31) is weakly tighter than that in (33), so (31) implies (33). Conversely, we will show that if (31) is violated, then (33) is also violated. Note that since

Z satisfies the Imbens and Angrist (1994) assumptions, $\Delta_k^Z(A)/\alpha_C$ is the LATE of D on the compound outcome $1\{Y,M=m_k\}$, i.e.

$$\frac{\Delta_k^Z(A)}{\alpha_C}=P(Y^{\text{tot}}(1)\in A,M(1)=m_k\,|\,C^z=1)-P(Y^{\text{tot}}(0)\in A,M(0)=m_k\,|\,C^z=1).$$

It follows that

$$\begin{aligned}\sup_A\Delta_k^Z(A)&=\alpha_C\sup_A\left\{P(Y^{\text{tot}}(1)\in A,M(1)=m_k\,|\,C^z=1)-P(Y^{\text{tot}}(0)\in A,M(0)=m_k\,|\,C^z=1)\right\}\\&\leqslant\alpha_C P(M(1)=m_k\,|\,C^z=1)\\&=E[D\cdot 1\{M=m_k\}\,|\,Z=1]-E[D\cdot 1\{M=m_k\}\,|\,Z=0]\end{aligned}\tag{35}$$

where the inequality uses the fact that $P(Y^{\text{tot}}(1)\in A,M(1)=m_k\,|\,C^z=1)\leqslant P(M(1)=m_k\,|\,C^z=1)$; and the final equality again uses the fact that Z satisfies the Imbens and Angrist (1994) assumptions to obtain that

$$P(M(1)=m_k\,|\,C^z=1)=\frac{E[D\cdot 1\{M=m_k\}\,|\,Z=1]-E[D\cdot 1\{M=m_k\}\,|\,Z=0]}{\alpha_C}.$$

It follows from (35) that (31) can be violated only if there exists some k such that

$$\sup_A\Delta_k(A)>P(M\geqslant m_k\,|\,Z=1)-P(M\geqslant m_k\,|\,Z=0),$$

in which case (33) is also violated. This completes the proof that the testable implications in Corollary 3.1 and Corollary 5.1 are equivalent under monotonicity of $M(d)$.

This equivalence breaks down if one does not impose monotonicity of $M(d)$. Intuitively, if we do not impose monotonicity, then by ignoring D we lose information about the type shares θ among instrument-compliers. A simple example is as follows. Suppose there are two groups in the population occurring with equal probability. The first group are instrument-compliers $(D(1)=1,D(0)=0)$ and always-takers with respect to M $(M(1)=M(0)=1)$ with $Y(d,m)=d$. Since $Y(d,m)$ depends on d, the sharp null is violated. The second group are instrument never-takers $(D(1)=D(0)=0)$ and never-takers with respect to M $(M(1)=M(0)=0)$ with $Y(d,m)=0$. Then it is straightforward to verify that (30) is violated whereas (32) is not. The reason for this is that the type shares using the potential outcomes are point-identified for instrument-compliers since $P(M(1)=1\,|\,C^z=1)=P(M(0)=1\,|\,C^z=1)=1$, and hence Θ^*_{I,C^z} is a singleton with unique element θ^* such that $\theta^*_{11}=1$ and $\theta^*_{lk}=0$ for $lk\neq 11$. Thus (30) reduces to the restriction that $\sup_A\Delta_k^Z(A)=0$, which is violated for $k=1$. On the other hand, since $P(M=1\,|\,Z=1)=P(M=1\,|\,Z=0)=0.5$, the identified set Θ_I based on Corollary 3.1 includes $\tilde{\theta}$ such that $\tilde{\theta}_{01}=\tilde{\theta}_{10}=0.5$ and $\tilde{\theta}_{kk}=0$ for all k, in which case there are no always-takers and thus (32) is trivially satisfied.

C Additional Results on Discretizing Continuous Outcomes

Connection to conditional moment inequalities. As discussed in Remark 9, implementing the proposed test after discretizing the outcome variable continues to yield a valid inference procedure under the sharp null. A drawback is that the resulting testable implications are potentially no longer sharp. Discretizing the outcome is analogous to the selection of instrument functions in tests of conditional moment inequalities, as in Andrews and Shi (2013). Indeed, in the simplest setting where M is binary and monotonicity holds, Mourifié and Wan (2017) show that the sharp testable implications can be reformulated as conditional moment inequalities of the form $E[g(M,D)\,|\,Y] \geqslant 0$, for g a function of M,D. Andrews and Shi (2013) propose tests of inequalities of the form $E[g(M,D)\,|\,Y] \geqslant 0$ that are based on unconditional inequalities of the form $E[h(Y)\cdot g(M,D)] \geqslant 0$, where h is a non-negative function of Y. Their recommended approach is to use hyper-cubes for h, which for a one-dimensional Y (as in our setting) corresponds to indicators for Y lying in particular intervals. Thus, applying the Andrews and Shi (2013) approach in this setting is equivalent to testing the Mourifié and Wan (2017) implications with a discretized outcome.[33] The choice of instrument functions for conditional moment inequalities is known to be a theoretically challenging question. As a result, existing recommendations are often heuristic in nature, motivated by simulation evidence.[34] For instance, based on Monte Carlo simulations, Andrews and Shi (2013) suggest choosing instrument functions such that the expected number of observations per cell lies between 10 and 20.

Monte Carlos. In a similar spirit, as mentioned in the main text, we conduct Monte Carlo simulations calibrated to Baranov et al. (2020) using 2 or 10 bins instead of the 5 bins used in the results reported in the main text (Appendix Tables 1 and 2). In Appendix Table 3, we report how the number of independent observations per cell varies across these specifications. Since the observations are clustered, rather than counting the raw observations, we count the number of independent clusters that have at least one observation in a given cell. Here, a cell corresponds to a support point of the vector (D,M,Y^{disc}), where Y^{disc} denotes the discretized outcome variable.

[33] In the simple case of binary M and monotonicity, one could apply other tests for conditional moment inequalities that are not analogous to discretizing the outcome, but which typically rely on other tuning parameters. For example, Mourifié and Wan (2017) suggest an approach based on Chernozhukov, Lee and Rosen (2013)'s test of conditional moment inequalities, which requires the choice of a bandwidth and kernel for non-parametric mean estimation. Whether such approaches could be extended to our more general setting with multi-valued M or relaxations of monotonicity strikes us an interesting question for future work.

[34] There are some formal results on *rate-optimal* choices of test statistic for conditional moment inequalities (Armstrong, 2014; Chetverikov, 2018), although these results do not appear to immediately apply to our general setting in which there are nuisance parameters (and thus we are interested in subvector inference).

For each simulated dataset, we compute the number of independent clusters per cell and record the median cell count. We then report the average of these median cell counts across simulation replications. See Section 4.1 for discussion of these results.

Appendix Table 1: Simulation results for Baranov et al. (2020) with binary M and different discretizations of the outcome

Panel A: Baranov et al, 40 clusters, 2 bins

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.086	0.078	0.050	0.136	0.126
t=0.5	0.134	0.264	0.256	0.064	0.314	0.280
t=1	0.283	0.828	0.822	0.422	0.844	0.830

Panel B: Baranov et al, 80 clusters, 2 bins

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.046	0.040	0.040	0.098	0.090
t=0.5	0.134	0.444	0.430	0.160	0.456	0.434
t=1	0.283	0.978	0.976	0.846	0.976	0.976

Panel C: Baranov et al, 200 clusters, 2 bins

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.052	0.044	0.030	0.082	0.078
t=0.5	0.134	0.822	0.816	0.618	0.818	0.796
t=1	0.283	1	1	1	1	1

Panel D: Baranov et al, 40 clusters, 10 bins

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.072	0.188	0.050	0.324	0.262
t=0.5	0.134	0.164	0.246	0.064	0.340	0.308
t=1	0.283	0.530	0.658	0.422	0.774	0.720

Panel E: Baranov et al, 80 clusters, 10 bins

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.052	0.086	0.040	0.208	0.158
t=0.5	0.134	0.272	0.314	0.160	0.436	0.368
t=1	0.283	0.798	0.924	0.846	0.960	0.942

Panel F: Baranov et al, 200 clusters, 10 bins

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.042	0.048	0.030	0.122	0.100
t=0.5	0.134	0.636	0.742	0.618	0.804	0.754
t=1	0.283	0.998	1	1	1	1

Notes: This table show simulation results analogous to Panels B-D of Table 1, with 2 and 10 bins used for discretizing the outcome variable. The first column shows the value of t, which determines the distance from the null, as described in the main text. The second column shows the lower-bound on the fraction of always-takers affected by treatment, $\bar{\nu}$. The remaining columns contain the rejection probabilities for each of the inference methods considered. Panels A-C use 2 bins to discretize the outcome variable and Panels D-F use 10 bins. Since Kitagawa (2015) does not require a discrete outcome variable, we use the outcome variable as-is when running this test (hence the results for K do not depend on the number of bins). Rejection probabilities are computed over 500 simulation draws, under a 5% significance level.

Appendix Table 2: Simulation results for Baranov et al. (2020) with non-binary M and different discretizations of the outcome

Panel A: Baranov et al, 40 clusters, 2 bins

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.056	0.092	0.150	0.112
t=0.5	0.119	0.092	0.206	0.356	0.326
t=1	0.255	0.290	0.856	0.944	0.922

Panel B: Baranov et al, 80 clusters, 2 bins

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.054	0.058	0.146	0.110
t=0.5	0.119	0.110	0.392	0.546	0.514
t=1	0.255	0.288	0.986	0.998	0.998

Panel C: Baranov et al, 200 clusters, 2 bins

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.042	0.048	0.100	0.076
t=0.5	0.119	0.104	0.792	0.892	0.860
t=1	0.255	0.422	1	1	1

Panel D: Baranov et al, 40 clusters, 10 bins

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.038	0.102	0.386	0.264
t=0.5	0.119	0.036	0.256	0.556	0.464
t=1	0.255	0.126	0.818	0.960	0.932

Panel E: Baranov et al, 80 clusters, 10 bins

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.048	0.032	0.282	0.176
t=0.5	0.119	0.050	0.238	0.650	0.566
t=1	0.255	0.134	0.986	0.998	0.998

Panel F: Baranov et al, 200 clusters, 10 bins

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.048	0.006	0.182	0.094
t=0.5	0.119	0.068	0.464	0.936	0.894
t=1	0.255	0.264	1	1	1

Notes: This table show simulation results analogous to Table 2, with 2 and 10 bins used for discretizing the outcome variable. The first column shows the value of t, which determines the distance from the null, as described in the main text. The second column shows the lower-bound on the fraction of always-takers affected by treatment, $\bar{\nu}$. The remaining columns contain the rejection probabilities for each of the inference methods considered. Panels A-C use 2 bins to discretize the outcome variable and Panels D-F use 10 bins. Rejection probabilities are computed over 500 simulation draws, under a 5% significance level.

Appendix Table 3: Average median cell count for DGPs calibrated to Baranov et al. (2020)

Panel A: Baranov et al, 40 clusters

	Binary M			Non-binary M		
	2 bins	5 bins	10 bins	2 bins	5 bins	10 bins
t = 0.0	19.976	13.999	9.751	15.236	8.943	5.904
t = 0.5	19.729	13.535	9.546	15.362	8.639	5.625
t = 1.0	19.584	13.245	9.299	15.654	8.832	5.417

Panel B: Baranov et al, 80 clusters

	Binary M			Non-binary M		
	2 bins	5 bins	10 bins	2 bins	5 bins	10 bins
t = 0.0	39.978	28.086	18.980	30.105	17.461	11.115
t = 0.5	39.298	26.851	18.923	30.562	16.720	10.916
t = 1.0	38.797	26.194	17.780	31.762	17.485	10.327

Panel C: Baranov et al, 200 clusters

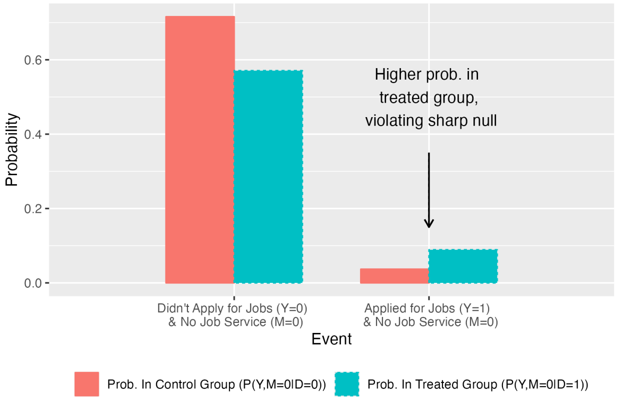
	Binary M			Non-binary M		
	2 bins	5 bins	10 bins	2 bins	5 bins	10 bins
t = 0.0	99.985	70.007	46.312	74.170	42.033	27.104
t = 0.5	98.084	66.870	47.000	75.869	38.988	27.176
t = 1.0	95.952	65.184	43.070	80.675	43.885	24.863

Notes: This table reports the median number of independent clusters per cell, averaged over 500 simulation replications. A cell corresponds to each support point of (D,M,Y^{disc}), where Y^{disc} is the discretized (to either 2, 5 or 10 bins) version of the outcome Y. The first three columns are calculated from the DGPs calibrated to Baranov et al. with binary M (i.e., DGPs considered in Table 1 Panels B-D and Appendix Table 1) and the last three columns are calculated from the DGPs calibrated to Baranov et al. with non-binary M (i.e., DGPs considered in Table 2 Panels B-D and Appendix Table 2).

D Additional Empirical Results

Alternative sample for Bursztyn et al. (2020). In our application to Bursztyn et al. (2020) in the main text, we restrict attention to the 75 percent of men who under-estimate other men's openness at baseline, which increases the plausibility of the monotonicity assumption. We now present analogous results using the full sample, which are similar. Appendix Figure 1 is analogous to Figure 1 but using the full sample, with similar qualitative patterns. The estimated lower bound on the fraction of never-takers affected, imposing monotonicity, is 8 percent, and bounds for the average effect for never-takers are 0.08 to 0.13. The lower bound on the fraction affected remains non-zero allowing for up to 5 percent of the population to be defiers.

Appendix Figure 1: Illustration of Testable Implications in Bursztyn et al. (2020) Using Full Sample



Note: This figure is analogous to Figure 1 except it uses the full sample rather than restricting to men who initially underestimate others' beliefs.

Alternative tests. In the main text, we report statistical tests of the sharp null using CS, using a discretization with 5 bins for the Baranov et al. (2020) application. Appendix Table 4 presents results for the Baranov et al. (2020) application alternatively using either 2 bins or 10 bins, with qualitatively similar conclusions. Appendix Table 5 reports test results using the tests of ARP and FSST instead of CS (using 5 bins for the Baranov et al. (2020) application).[35] The qualitative pattern across the tests is similar. One notable difference is that we do not reject the null for the relationship-quality mechanism in Baranov et al. (2020) using ARP, although this is perhaps unsurprising given the low power of ARP in simulations calibrated to this mechanism.

[35] Recall that the reported p-value is the smallest value of α for which the test rejects. Since ARP uses a two-stage procedure, it is difficult to analytically compute the p-value. We therefore compute the test for α values on a grid with interval-length 0.01 between 0.01 and 0.1 and interval-length 0.1 between 0.15 and 0.95, and report the smallest grid point at which the test rejects.

Appendix Table 4: p-values for tests for the sharp null in Baranov et al. (2020) using alternative bin choices

	Number of bins		
Mediator	2	5	10
Grandmother	0.003	0.023	0.065
Relationship	0.005	0.028	0.001
Grandmother + Relationship	0.198	0.654	0.999

Appendix Table 5: p-values for tests for the sharp null using alternative procedures

Application	M	CS	ARP	FSSTdd	FSSTndd
Bursztyn et al (main sample)	Job-search Sign-up	0.020	0.030	0.018	0.020
Bursztyn et al (full sample)	Job-search Sign-up	0.019	0.020	0.021	0.022
Baranov et al	Grandmother	0.023	0.030	0.026	0.047
Baranov et al	Relationship	0.028	0.650	0.037	0.049
Baranov et al	Grandmother + Relationship	0.654	0.550	0.115	0.256