Difference-in-Differences for Ordinal Outcomes: Application to the Effect of Mass Shootings on Attitudes towards Gun Control[*]

Soichiro Yamauchi[†]

This version: March 28, 2020
First draft: October 29, 2019

Abstract

The difference-in-differences (DID) design is widely used in observational studies to estimate the causal effect of a treatment when repeated observations over time are available. Yet, almost all existing methods assume linearity in the potential outcome (*parallel trends* assumption) and target the additive effect. In social science research, however, many outcomes of interest are measured on an ordinal scale. This makes the linearity assumption inappropriate because the difference between two ordinal potential outcomes is not well defined. In this paper, I propose a method to draw causal inferences for ordinal outcomes under the DID design. Unlike existing methods, the proposed method utilizes the latent variable framework to handle the non-numeric nature of the outcome, enabling identification and estimation of causal effects based on the assumption on the quantile of the latent continuous variable. The paper also proposes an equivalence-based test to assess the plausibility of the key identification assumption when additional pre-treatment periods are available. The proposed method is applied to a study estimating the causal effect of mass shootings on the public's support for gun control. I find that the effect is concentrated on left-leaning respondents who experienced the shooting for the first time in more than a decade.

Keywords: Difference-in-differences, gun control, ordinal outcome, panel data

[*] I am very grateful to Matt Blackwell, Kosuke Imai, Ikuma Ogura, members of Imai research group at Harvard (Soubhik Barari, Jake Brown, Naoki Egami, Shusei Eshima, Max Goplerud, June Hwang, Connor Jerzak, Shiro Kuriwaki, Santiago Olivella, Sun Young Park, Casey Petroff, Avery Schmidt, Sooahn Shin, Tyler Simko and Diana M. Stanescu) and participants of G3 Mini-Conference for comments and suggestions. I also thank Shusei Eshima, Shiro Kuriwaki, and Shun Yamaya for insightful discussions. R package orddid is available for implementing the proposed methodology at https://github.com/soichiroy/orddid.

[†] Graduate student, Department of Government and Institute for Quantitative Social Science, Harvard University. Email: syamauchi@g.harvard.edu. URL: https://soichiroy.github.io/.

1 Introduction

The difference-in-differences (DID) design is widely used in observational studies with repeated observations over time (Card and Krueger, 1994; Angrist and Pischke, 2008; Lechner et al., 2011). It allows scholars to identify the causal effect accounting for time-invariant unobserved confounders. Although significant progress has been made to improve the original DID design in recent years (e.g., Abadie, 2005; Athey and Imbens, 2006; Qin and Zhang, 2008; Lee, 2016; Arkhangelsky et al., 2018; Callaway and Sant'Anna, 2018; Li, 2019; Lu et al., 2019), most of the existing methods attempt to identify and estimate the average treatment effect on the treated (ATT) under the so-called "parallel trends" assumption (Abadie, 2005). One common definition of ATT is additive, defined as a difference between two potential outcomes averaging over the treated units. This additive effect is well defined when the outcome is continuous where the difference between two potential outcomes can be meaningfully interpreted. In addition, the "parallel trends" assumption imposes linearity on the trends of the outcome (Athey and Imbens, 2006). Therefore, the assumption and the estimand can be meaningful when the outcome is continuously distributed.

In social science research, however, many outcomes of interest are measured on an ordinal scale. For example, in survey research, most of the measurements are made on an ordinal scale (e.g., Likert, 1932). When the outcome is measured on an ordinal scale, the usual definition of the treatment effect as the difference between two potential outcomes is not well defined (e.g., Volfovsky et al., 2015; Lu et al., 2018), unless strong assumptions about the scale are imposed. Furthermore, the standard identification assumption for DID that utilizes linearity also fails in this context. This implies that the standard DID cannot be directly used for ordinal outcomes.

With a dearth of methodologies tailored for analyzing ordinal outcomes in the DID setting, scholars often dichotomize the ordinal outcome with some threshold. Although it appears that this transformation enables scholars to adopt the standard DID method to estimate casual effects, this strategy is not robust to different transformations (i.e., different choices of the threshold). Specifically, due to the non-linear nature of the ordinal outcome, the validity of identification assumption (e.g., "parallel trends" assumption) under one transformation does not guarantee the validity of another transformation. This is problematic because often times scholars do not have substantive justification on which transformation should be employed.

In this paper, I develop a methodology for estimating causal effects for ordinal outcomes with repeated observations. Instead of assuming linearity on the actual outcome, I utilize the latent variable formation often used in the discrete choice models. Because the assumptions are imposed on the entire distribution of the latent variables, the proposed method does not require a transformation of the outcome variable. Furthermore, the method enables researchers to estimate interpretable causal effects, defined as a difference between two probabilities, under a single set of assumptions. I also propose a diagnostic tool when scholars have data from two pre-treatment periods. As in the standard DID for the continuous outcome, where scholars can check if the pre-treatment trends are parallel, this diagnostic test allows researchers to formally confirm whether the assumption holds at least during the pre-treatment periods. Specifically, it is a testing procedure based on a functional equivalence test to assess the plausibility of one of the key assumptions.

The method of this paper is closely connected to the literature on non-linear DID (e.g., Athey and Imbens, 2006; Sofer et al., 2016; Callaway et al., 2018; Glynn and Ichino, 2019). In particular, Athey and Imbens (2006) consider an extension of their method to the binary and the count outcomes, but they do not consider a case of ordinal outcomes. Most importantly, because they impose a minimum restriction on the potential outcome, their method does not provide point identification even for an additive effect and the bound can be uninformative. Instead, I impose a stronger assumption for the sake of identifying the non-additive causal effect, which provides informative estimates of the causal effect for researchers.

The proposed methodology is used to revisit a recent debate on the effect of mass shootings on attitudes towards gun control regulations (Barney and Schaffner, 2019; Hartman and Newman, 2019; Newman and Hartman, 2019). In their original and the follow-up study, Hartman and Newman argue that a mass shooting increases people's support for stricter gun controls regardless of their partisanship. They also argue that the effect is conditional on the context around their life such as the safety of their neighborhood. On the other hand, Barney and Schaffner argue that there is no strong evidence to support the claim of Hartman and Newman. They also find a polarizing effect of mass shootings where Democrats become more supportive of gun control while Republicans become less supportive of gun control.

In Section 4, I analyze the data from the motivating empirical study using the proposed method. Using the two-wave panel data, I find that mass shootings have an effect on those who experience mass shootings for the first time in a decade in that they form a stronger opinion about gun control regulations, while the result suggests little evidence for the pro-regulation swing. I also find that the effect is concentrated among Democrats including those who weakly identify themselves as Democrat. However, the effects among Republicans are not statistically distinguishable from zero; thus I find little evidence to support the polarizing effect of mass shootings. Using three-wave panel data, which provides pre-treatment time periods, I test the plausibility of one of the assumptions made in the analysis and find that the data is consistent with the assumption. Reanalysis of the three-wave panel, however, finds little evidence to support the claim that mass shootings have any effect on the support for gun control.

The rest of the paper is organized as follows. Section 2 introduces the motivating application of the method. In Section 3, the methodology is introduced where I discuss identification assumptions and estimation strategy. In Section 4, I apply the proposed method to the data described in Section 2. Finally, I offer concluding remarks in Section 5.

2 The Effect of Mass Shootings on Public Support for Gun Control

2.1 The debate on the effect of mass shootings

This section describes observational studies that investigate an important issue using an ordinal outcome measured over time. Newman and Hartman (2019) and the follow-up studies (Barney and Schaffner, 2019; Hartman and Newman, 2019) study the relationship between experiencing mass shootings and the attitude to gun control. These studies use survey data with two-wave and three-wave panel to investigate whether living in close proximity to mass public shootings would have causal impact on people's attitude to gun control. Respondents to the survey are considered as "treated" if at least one mass shooting occurs within 100 miles from their address (measured at the zip code level). To measure the attitude to gun control, the authors used a response to the following survey question in the Cooperative Congressional Election Study (CCES):

> In general, do you feel that laws covering the sale of firearms should be made more strict, less strict, or kept as they are?
>
> (0) Less Strict; (1) Kept As They Are; (2) More Strict.

The original studies utilize variations in treatment assignment over time to isolate the effect of mass shootings from the time trends and location effects. Based on the analysis, Hartman and Newman find that living in near proximity to mass public shootings affects people to support stricter regulations on gun sales (Newman and Hartman, 2019; Hartman and Newman, 2019). They also report that the effect does not vary by respondents' party affiliation. In the follow-up study conducted by Barney and Schaffner (2019), authors correct data and conduct additional analyses. They conclude that the effect varies by which party people affiliate with and that there is a polarizing effect of mass shootings in that Democrats become more supportive of stricter gun control, while Republican become less supportive of regulations.

Throughout the debate, the authors utilize a variety of methodologies such as ordered probit model with random effects and linear fixed effect model to estimate the impact of mass shootings on the attitude (see Table 1 in Appendix E). Although "difference-in-differences" is mentioned in these studies, discussions about the quantities of interest and assumptions required for the identification of those quantities are missing from the debate. Without assessing the assumptions explicitly, it is simply challenging to draw a conclusion if mass shootings have any effect on people's attitude. Thus, the goal of the paper is to fill this gap by proposing a methodology that enables researches to assess assumptions and reliably estimate causal effects.

2.2 The challenge

For a causal effect to be reliably estimated, assumptions have to be imposed and assessed. In the following, I demonstrate that even if researchers are aware of identification assumptions, the current practice is not well suited for the ordinal outcome. Suppose that, following the common practice, researchers dichotomize an ordinal outcome into a binary outcome by specifying some threshold. The standard DID analysis is then applied on this transformed outcome. In our running example, there are two possible ways to transform the original outcome into a binary variable. One way is to code `more-strict` category as one and `kept-as-they-are` and `less-strict` as zero and the other way is to treat `more-strict` and `kept-as-they-are` as one and `less-strict` as zero.

After transforming the outcome, scholars can check if the pre-treatment trends are parallel or not for this new binary variable. Inspecting pre-treatment trends is a routine often used in empirical studies to justify the use of the DID design (Angrist and Pischke, 2008). Figure 1 shows trends for each transformed outcome using the three-wave panel of the survey where I subset respondents who are not treated until 2012. The panel on the left shows the first type of transformation where only `more-strict` category is coded as 1 (denoted by Y on the y-axis). We can see that the pre-treatment trends between the treatment group (blue) and the control group (gray) appears to be parallel (denoted by D on the y-axis.). Thus if a researcher transforms the original variable in this way, she might conclude that the DID

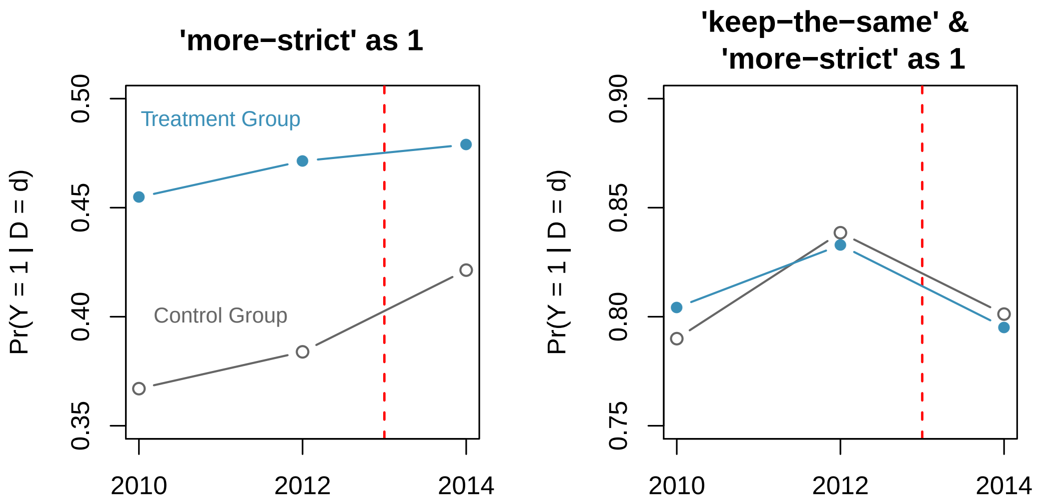

Figure 1: Visual assessment of parallel trends assumption for the three-wave panel from the survey data. Respondents who are not treated until 2012 are used for generating this plot. The lines with solid circles (blue) show trends for the treated group and the lines with hollow circles (gray) show trends for the control group. Vertical dashed lines (red) show the timing of the treatment. Although the left panel appears to show that pre-treatment trends are parallel between the treatment and the control group, the right panel suggests that pre-treatment trends are not parallel.

design is suitable for analyzing the data. The panel on the right shows the pre-treatment trends for the second type of transformation. In this case, however, the pre-treatment trends do not seem to be parallel, rather two trends are crossing during the pre-treatment periods.

Note that this observation is not specific to this application. In Appendix C, I demonstrate that it is trivial to construct an example that satisfying parallel trends in one transformation does not imply the parallel trends in another transformation.

It is often unclear *ex ante* which threshold should be chosen from a substantive point of view. Therefore, it is unfortunate that the validity of the design seems to depend on how the variable is transformed. Although the running example only has three categories, the problem exacerbates when scholars need to analyze an outcome that has a larger number of categories.

3 The Proposed Methodology

3.1 The setup

Let $Y_{it} \in \{0, \dots, J-1\} \equiv \mathcal{J}$ denote the observed outcome measured on an ordinal scale with J categories ($J \geq 3$) for unit $i \in \{1, \dots, n\}$ and time $t \in \{0, 1\}$. The binary treatment, denoted by $D_i \in \{0, 1\}$, is assigned after Y_{i0} is observed but before time $t = 1$. We use the potential outcome notation to denote the counterfactual outcome, $Y_{it}(d)$ for $d \in \{0, 1\}$. For example, $Y_{i1}(0)$ is an attitude toward gun control regulations that would realize in the post-period if a respondent did not experience a mass shooting (i.e., the control condition).

In many applications, scholars are interested in estimating the distributional treatment effect. In this paper, I focus on the treatment effect on the treated. Specifically, the effect ζ_j is defined as the difference in probabilities of choosing category j under two conditions,

$$\zeta_j = \mathbb{P}(Y_{i1}(1) = j \mid D_i = 1) - \mathbb{P}(Y_{i1}(0) = j \mid D_i = 1). \tag{3.1}$$

for $j \in \mathcal{J}$. In our application, ζ_2 is the difference in probabilities that those treated prefer more strict gun control between the treated and the control conditions. Thus, observing $\zeta_2 > 0$ implies that the mass shootings make people prefer stricter policies on gun control. Similarly, ζ_0 is the effect of the treatment on `less-strict` category and $\zeta_0 > 0$ implies that incidents turn people to prefer less strict regulations.

When the number of categories is large, it is sometimes useful to estimate the cumulative effect Δ_j, which is defined as a difference in probabilities of choosing j or larger categories under the two conditions,

$$\Delta_j = \mathbb{P}(Y_{i1}(1) \geq j \mid D_i = 1) - \mathbb{P}(Y_{i1}(0) \geq j \mid D_i = 1) \tag{3.2}$$

for $j \in \mathcal{J} \backslash \{0\}$. Note that $\Delta_j = \sum_{\ell=j}^{J-1} \zeta_\ell$ by construction, and thus it is sufficient to consider the identification of ζ_j.

The cumulative effect is also useful to connect the approach that dichotomizes ordinal outcomes to the proposed method. From the above definition, we can see that the standard DID based on the dichotomized outcome at threshold j identifies Δ_j. This is because $\Pr(Y_{i1}(d) \geq j \mid D_i = 1) = \mathbb{E}[\mathbf{1}\{Y_{i1}(d) \geq j\} \mid D_i = 1] \equiv \mathbb{E}[\widetilde{Y}_{i1}(d) \mid D_i = 1]$ where $\widetilde{Y}_{i1}(d)$ is the dichotomized potential outcome with threshold j. This means that the standard DID applied to the dichotomized outcome can estimate only one of $J - 1$ possible quantities of interest. Furthermore, if one wishes to estimate all possible Δ_j's by changing the threshold, it requires $J-1$ distinct identification assumptions. As we saw in our motivating example, however, satisfying the parallel trends assumption for Δ_j does not necessarily imply that the assumption for $\Delta_{j'}$ is satisfied.

3.2 Identification

Typically, we do not have a good sense of which estimand is best suited for answering the substantive question. Therefore, it is natural that we attempt to identify and estimate ζ_j for all $j \in \mathcal{J}$ from the observed data. The goal of this section is to establish the identification of $\boldsymbol{\zeta} = (\zeta_0, \dots, \zeta_{J-1})^\top$ with a single set of assumptions.

To compute the quantity defined in Equation (3.1), we need the marginal distribution of $Y_{i1}(1)$ and $Y_{i1}(0)$ for the treated. While we observe $Y_{i1}(1)$ for $D_i = 1$ because $Y_{i1} = D_i Y_{i1}(1) + (1-D_i) Y_{i1}(0)$, we need to impose additional assumptions to identify the distribution of $Y_{i1}(0)$ for $D_i = 1$. Following Athey and Imbens (2006), I omit the subscript i for units and denote $Y_{dt} \sim Y_{it}(0) \mid D_i = d$ where $A \sim B$ indicates A and B are equivalent in distribution. Y_{dt} denotes the potential outcome under the control condition at time t for group defined by $D_i = d$. While we observe Y_{00}, Y_{01} and Y_{10}, the counterfactual outcome $Y_{11} \sim Y_{i1}(0) \mid D_i = 1$ is what we do not observe in the data. In our example, Y_{11} is the potential attitude to gun control that we would have observed if those respondents who have experienced mass shootings would have not been exposed to the event.

I first impose a structure on the potential outcome. Specifically, I assume that the observed categorical outcome follows the index model, which means that there is a latent variable behind Y_{dt} and that the categorical outcome is defined by a simple thresholding rule on the latent variable.

Assumption 1 (Index model)**.** Assume that the potential outcomes follow the index model such that there exists a latent variable $Y^*_{dt} \in \mathbb{R}$ and

$$Y_{dt} = \begin{cases} 0 & \text{if} \quad \kappa_0 \leq Y^*_{dt} < \kappa_1 \\ j & \text{if} \quad \kappa_j \leq Y^*_{dt} < \kappa_{j+1} \\ J-1 & \text{if} \quad \kappa_{J-1} \leq Y^*_{dt} \leq \kappa_J \end{cases} \tag{3.3}$$

where $\{\kappa_j\}_{j=0}^{J}$ are a set of cutoffs with $\kappa_0 = -\infty$ and $\kappa_J = \infty$.

Assumption 1 says that the potential outcome defined on an ordinal scale Y_{dt} is a function of another potential outcome defined on a continuous space Y^*_{dt}. In the application, Y^*_{dt} can be considered as the underlying intensity of one's attitude toward gun control policies where larger value of Y^*_{dt} corresponds to a support for stricter gun control. The assumption allows us to handle the outcome on a continuous space through Y^*_{dt} instead of directly working on a discrete space. Note that κ_j's are constants assumed to be fixed and they do not depend on group (d) nor time (t).

Different from the additive effect, the distributional treatment effect ζ_j requires that the entire marginal distribution of the potential outcome is identified. For that, I further impose a distributional assumption on Y^*_{dt} in Assumption 2.

Assumption 2 (Location-scale family assumption)**.** Let U denote a continuously distributed random variable with mean 0 and variance 1 that belongs to a parametric family. We assume that Y^*_{dt} belongs to the location-scale family, that is, it can be written as

$$Y^*_{dt} \sim \mu_{dt} + \sigma_{dt} U \tag{3.4}$$

where μ_{dt} is the location and σ_{dt} is the scale parameter.

Assumption 2 specifies the distribution of the latent utilities. It assumes that each marginal distribution belongs to the location-scale family distribution with time and group specific location and scale parameter. This implies that the distribution of the potential outcomes are different up to mean and the scale. Note that the joint distribution of the latent utilities are left unspecified, so units can have correlated latent utilities over time. Although this is a parametric assumption (i.e., the distribution of U should be known), the location-scale family encompasses a large class of parametric distributions (e.g., the normal distribution, the logistic distribution or the t-distribution, etc).

Finally, I impose a structure on the relationship between latent variables Y^*_{dt}. This allows us to map what we observe in the control group over time to what would have happened to the treated group if it was not treated. I first start with a restrictive assumption that is similar to the standard DID design. It is possible to assume that the parallel trends hold on the latent outcome, that is,

$$\mathbb{E}[Y^*_{i1} \mid D_i = 1] - \mathbb{E}[Y^*_{i0} \mid D_i = 1] = \mathbb{E}[Y^*_{i1} \mid D_i = 0] - \mathbb{E}[Y^*_{i0} \mid D_i = 0]. \tag{3.5}$$

Then, the mean of the counterfactual latent outcome Y_{11}^* is uniquely identified as

$$\mu_{11} = \mu_{10} + \mu_{01} - \mu_{00}.$$

However, this approach is restrictive, because it requires an additional assumption that the variance is constant across time and groups, that is, $\sigma_{dt} = \sigma$ for all d and t; otherwise we cannot identify the entire distribution of the latent outcome Y_{11}^*. This constant variance assumption is strong because it only allows the unidirectional change of choice probabilities.

Therefore, I impose a different assumption from the standard parallel trends assumption. Instead assuming the mean shift, the assumption is imposed on the entire distributions, which is originally introduced by Athey and Imbens (2006) (also see Sofer et al., 2016). Specifically, I assume that the shift in the distribution across time are constant between the treatment and the control groups. Figure 2 graphically illustrates the assumption. The key part of this assumption is that the vertical arrows in the two graphs should be the same length. In other words, $q_d(v) - v$ captures the trend in the distribution (i.e., how much Y_{dt}^* "shifts" between $t = 0$ and $t = 1$) and the assumption says that the "shift" is identical across two groups. This means that for each choice of v, the corresponding value of $q_d(v)$ ("shift") should be the same for $d = 0, 1$.

Assumption 3 formally states the assumption.

Assumption 3 (Distributional parallel trends (Athey and Imbens, 2006))**.** Let $F_{Y_{dt}^*}(y) = \mathbb{P}(Y_{dt}^* \leq y)$ be the cumulative distribution function (CDF) of Y_{dt}^* and define $q_d(v) = F_{Y_{d0}^*} \circ F_{Y_{d1}^*}^{-1}(v)$. Then, we assume that for all $v \in [0, 1]$,

$$q_1(v) = q_0(v) \tag{3.6}$$

Assumption 3 imposes a restriction on the relationship between the pre-treatment latent outcome Y_{10}^* and the counterfactual latent outcome Y_{11}^*, based on the relationship between two latent variables in the control group. Note that by construction $q_1(v) - q_0(v) = 0$ for $v = 0, 1$ because CDFs should agree at the end of the support, $\lim_{y \to \pm\infty} F_{Y_{d0}^*}(y) = \lim_{y \to \pm\infty} F_{Y_{d1}^*}(y)$.

Assumption 1, 2 and 3 identify the distribution of the counterfactual outcome. Proposition 1 presents the formal result.

Proposition 1 (Identification of the Counterfactual Distribution)**.** Under Assumption 1, 2,

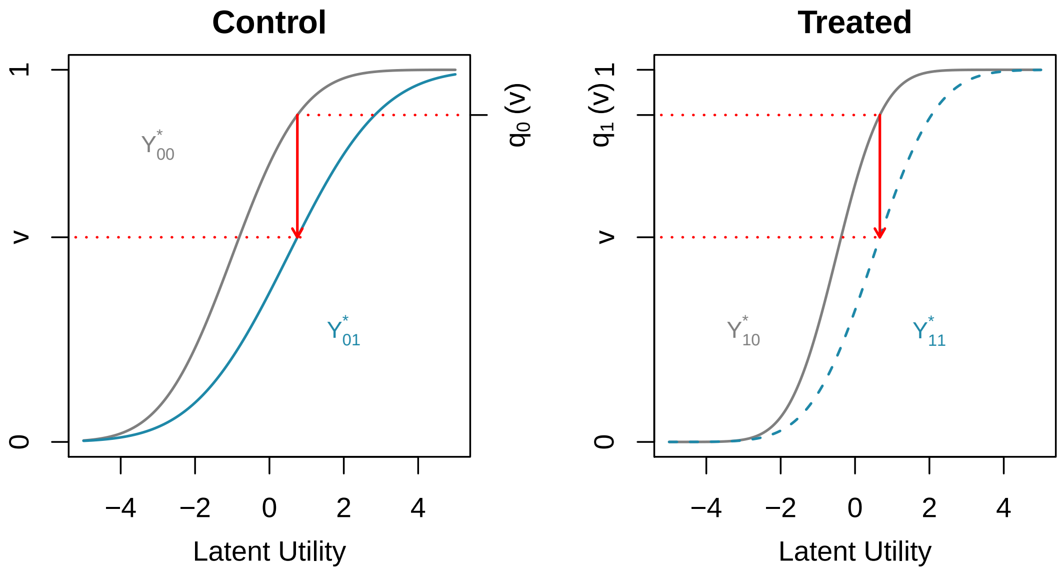

Figure 2: Graphical illustration of Assumption 3. Left (right): cumulative distribution functions of the latent utilities Y^*_{dt} under the control (treatment) condition. Blue (gray) lines indicate the distribution for time $t = 1$ ($t = 0$). Dashed line on the right panel is the distribution of counterfactual outcome $Y^*_{11} \sim Y^*_{i1}(0)|D_i = 1$. The key assumption is that the length of the vertical arrow (red) is the same between the two panels for all range of v. This allows us to recover the shape of the dashed line based on latent utility distributions for the observed outcomes (i.e., solid lines).

and 3, the distribution of the counterfactual latent utility Y^*_{11} is identified as

$$Y^*_{11} \sim \mu_{11} + \sigma_{11} U \tag{3.7}$$

where

$$\mu_{11} = \mu_{10} + \frac{\mu_{01} - \mu_{00}}{\sigma_{00}/\sigma_{10}} \quad \text{and} \quad \sigma_{11} = \frac{\sigma_{10}\sigma_{01}}{\sigma_{00}}.$$

And thus, the distribution of the potential outcome is identified as

$$\mathbb{P}(Y_{i1}(0) = j \mid D_i = 1) = F_U\left(\frac{\kappa_{j+1} - \mu_{11}}{\sigma_{11}}\right) - F_U\left(\frac{\kappa_j - \mu_{11}}{\sigma_{11}}\right)$$

for $j = 0, \ldots, J-1$, where $F_U(u) = \mathbb{P}(U \leq u)$ is the CDF of U.

A proof is in Appendix A. Proposition 1 says that the location and the scale of the counterfactual latent outcome Y^*_{11} are uniquely determined by parameters of observed outcomes. This implies that we can recover the distribution of the counterfactual outcome $Y_{11} \sim Y_{i1}(0)|D_i = 1$ (i.e., the potential outcome under the control condition for the treated unit at time $t = 1$) using parameters estimated from the observed data, Y_{00}, Y_{01} and Y_{10}. For example, if we assume that U follows the standard normal distribution, we have that Y_{11}^* follows the normal distribution with mean μ_{11} and variance σ_{11}^2.

3.3 Estimation

The identification result in the previous section provides a guidance on how we can estimate the causal effect from the observed data. Let $\boldsymbol{\theta}_{dt} = (\mu_{dt}, \sigma_{dt})^\top$ denote a vector of parameters that characterize the distribution of the latent utility Y_{dt}^*. I take a two-step approach to estimate causal quantity ζ_j defined in Equation 3.1 for all $j \in \mathcal{J}$. In the first step, I estimate parameters for observed outcomes, that is, $\boldsymbol{\theta}_{00}$, $\boldsymbol{\theta}_{01}$ and $\boldsymbol{\theta}_{10}$. Based on the estimate of these parameters, causal effects are estimated in the second step.

In this and the following section, I focus on a case where Y_{dt}^* follows the normal distribution, that is $U \sim \mathcal{N}(0, 1)$. Then, by Assumption 2, the observed outcomes Y_{00}, Y_{01} and Y_{10} follow the ordered probit model. Thus, parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{00}^\top, \boldsymbol{\theta}_{01}^\top, \boldsymbol{\theta}_{10}^\top, \boldsymbol{\kappa}^\top)^\top$ can be estimated via the maximum likelihood.

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\kappa}} \sum_{i=1}^{n} \sum_{t \in \{0,1\}} \sum_{j \in \mathcal{J}} \mathbf{1}\{Y_{it} = j, tD_i = 0\} \log \Big\{ \Phi[(\kappa_{j+1} - \mu_{D_i,t})/\sigma_{D_i,t}] - \Phi[(\kappa_j - \mu_{D_i,t})/\sigma_{D_i,t}] \Big\}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function that takes 1 if the argument is true and takes 0 otherwise, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Different from the standard ordered probit specification, I fix two cutoffs κ_1 and κ_2 (recall that $\kappa_0 = -\infty$ and $\kappa_J = \infty$). This allows us to estimate the variance component in addition to means (Lemma 1 in Appendix A; also see for example Jackman, 2009, Chapter 8). Note that the choice of κ is not consequential in that, the causal effect estimate $\widehat{\boldsymbol{\zeta}}$ is invariant to the choice of the cutoffs (Lemma 3 in Appendix A). This is because the identification assumption imposes a structure on the quantile scale, which is invariant to the scale of the latent variables, while different choices of cutoffs only affect the location and the scale (i.e., μ and σ) of the latent variables.

We then estimate the parameter for the counterfactual distribution $\boldsymbol{\theta}_{11} = (\mu_{11}, \sigma_{11})^\top$ by the plug-in estimator based on the first stage,

$$\widehat{\mu}_{11} = \widehat{\mu}_{10} + (\widehat{\mu}_{01} - \widehat{\mu}_{00})/(\widehat{\sigma}_{00}/\widehat{\sigma}_{10}), \quad \text{and} \quad \widehat{\sigma}_{11} = (\widehat{\sigma}_{10}\widehat{\sigma}_{01})/\widehat{\sigma}_{00}. \tag{3.8}$$

Since the causal effect is a function of $\boldsymbol{\theta}_{11}$, the estimator for the causal effect is therefore given by

$$\widehat{\zeta}_j = \frac{1}{n_1}\sum_{i=1}^{n} D_i \mathbf{1}\{Y_{i1} = j\} - \Big\{\Phi[(\kappa_{j+1} - \widehat{\mu}_{11})/\widehat{\sigma}_{11}] - \Phi[(\kappa_j - \widehat{\mu}_{11})/\widehat{\sigma}_{11}]\Big\} \tag{3.9}$$

where $n_1 = \sum_{i=1}^{n} D_i$, and then $\widehat{\Delta}_j = \sum_{\ell=j}^{J-1} \widehat{\zeta}_j$. Note that the first term of the right-hand side is a nonparametric estimator of $\mathbb{P}(Y_{i1}(1) = j \mid D_i = 1)$ because this quantity is identified from the data without any assumptions. The second term on the right-hand side is the counterfactual distribution identified by the assumptions (Proposition 1).

Lemma 4 in Appendix A establishes the $\sqrt{n}$ consistency of the estimator $\widehat{\boldsymbol{\zeta}}$, whose sampling variance can be derived using the delta method under the independence assumption. In practice, however, the block bootstrap can be used to estimate the variance when outcomes are correlated across time or due to clustering.

3.4 Assessing the distributional parallel trends assumption

In the standard DID design, additional pre-treatment periods provide an opportunity to assess the parallel trends assumption by checking the pre-treatment trends (Angrist and Pischke, 2008; Egami and Yamauchi, 2019). Although it is not a direct test of the assumption, observing the parallel trends in the pre-treatment periods suggests that the assumption is more likely to be plausible. With a similar logic, we can assess the validity of distributional parallel trends assumption (Assumption 3). Specifically, we would expect that if the distributional parallel trends holds for the pre-treatment periods, it is more reasonable to claim that the assumption holds in the post-period. Thus, we assess the validity of Assumption 3 by testing if the similar condition holds for the pre-treatment periods.

The proposed testing procedure Suppose that we now observe the outcome for three time periods, Y_{i0}, Y_{i1} and Y_{i2} where Y_{i2} is the post-treatment outcome and Y_{i0} and Y_{i1} are the pre-treatment outcomes. The treatment is administered after time $t = 1$ in this setup, and thus we have $Y_{it}(0) = Y_{it}^{\text{obs}}$ for $t = 0, 1$ regardless of the treatment status. This means that observed outcome before the treatment assignment is the same as the potential outcome under the control condition for both treatment and control groups.

Let $\tilde{q}_d(v) = \Phi(\mu_{d0}, \sigma_{d0}) \circ \Phi^{-1}(v : \mu_{d1}, \sigma_{d1})$ denote the pre-treatment along of $q_d(v)$ defined in Assumption 3, where I assume that $U \sim \mathcal{N}(0, 1)$. Recall that $q_d(v)$ captures shift of distributions over time evaluated at quantile v. The assumption requires that two functions are identical on the unit interval, that is, $q_1(v) = q_0(v)$ for all v. Therefore, we wish to statistically test if $\tilde{q}_1(v) = \tilde{q}_0(v)$ holds for all $v \in [0,1]$ using the data from the pre-treatment periods.

Intuitively, we can check the equivalence of two functions $\tilde{q}_1$ and $\tilde{q}_0$ by assessing the maximum deviation between two functions, $t_{\max} = \max_{v\in[0,1]} |\tilde{q}_1(v) - \tilde{q}_0(v)|$. If this metric is "small", we may conclude that $\tilde{q}_1 = \tilde{q}_0$. Formally, with some threshold $\delta > 0$, we wish to test the following hypotheses:

$$H_0\colon \max_{v\in[0,1]} |\tilde{q}_1(v) - \tilde{q}_0(v)| > \delta \quad \text{and} \quad H_1\colon \max_{v\in[0,1]} |\tilde{q}_1(v) - \tilde{q}_0(v)| \leq \delta$$

where H_0 says that two functions are not equivalent (i.e., large deviation). Rejecting the null implies that the data supports H_1 of equivalence which is what we want to demonstrate. For now, I assume that researchers know how to choose an appropriate value of δ based on substantive knowledge. I will discuss how to calibrate this equivalence threshold in the below. We can see that the null hypothesis can be written as a union of two hypotheses without absolute values, $H_0 = H_0^+ \cup H_0^-$ where

$$H_0^+\colon \max_{v\in[0,1]} \{\tilde{q}_1(v) - \tilde{q}_0(v)\} > \delta \quad \text{and} \quad H_0^-\colon \min_{v\in[0,1]} \{\tilde{q}_1(v) - \tilde{q}_0(v)\} < -\delta.$$

This decomposition implies that we can conduct two one-sided tests to determine if we reject the original null H_0 or not. In other words, we conclude that H_0 is false if we reject *both* H_0^+ and H_0^-.

Now, suppose that we construct a $100(1-\alpha)\%$ point-wise confidence interval $[\widehat{L}_{1-\alpha}(v), \widehat{U}_{1-\alpha}(v)]$ for $t(v) \equiv \tilde{q}_1(v) - \tilde{q}_0(v)$ at each v. The detail of how to construct the confidence interval is presented in Lemma 6 and 7 in Appendix A. Then, by the one-to-one relationship between the test and the confidence set, we reject H_0^+ if and only if the upper confidence interval is less than δ, that is

$$\text{reject } H_0^+ \text{ at } \alpha \text{ level} \iff \max_{v\in[0,1]} \widehat{U}_{1-\alpha}(v) < \delta.$$

By the similar argument, we reject H_0^- at α level if and only if $\min_{v\in[0,1]} \widehat{L}_{1-\alpha}(v) > -\delta$.

Proposition 2 shows that the proposed procedure is in fact asymptotically level α test, that is, it rejects the null of non-equivalence with probability less than α when the null is true.

Proposition 2 (Validity of the Testing Procedure)**.** For a given choice of the equivalence threshold δ and the level of a test α, the testing procedure asymptotically controls the type I error, that is, under the null $H_0\colon t_{\max} \geq \delta$,

$$\sup_{t\colon \delta \leq |t| < 1} \mathbb{P}\Big(\Big\{ \max_{v \in [0,1]} \widehat{U}_{1-\alpha}(v) < \delta \Big\} \cap \Big\{ \min_{v \in [0,1]} \widehat{L}_{1-\alpha}(v) \geq -\delta \Big\}\Big) \leq \alpha$$

as $n \to \infty$.

The above proposition shows that when the equivalence threshold is chosen such that the null is true (i.e., $t_{\max} \geq \delta$), then the probably to falsely reject the null (type I error) is less than α, for any value of t that is consistent with the null. In other word, the proposed testing procedure is statistically valid for any choice of the equivalence threshold under the null.

The above result suggests that we can also compute the p-value for this test by solving the rejection rule with respect to α,

$$\widehat{p} = \max\Big\{ \max_{v \in [0,1]} \widehat{p}_1(v),\ \max_{v \in [0,1]} \widehat{p}_2(v) \Big\},$$

where

$$\widehat{p}_1(v) = 1 - \Phi\left(\frac{\delta - \hat{t}(v)}{\sqrt{\mathrm{Var}(\hat{t}(v))/n}} \right) \quad \text{and} \quad \widehat{p}_2(v) = 1 - \Phi\left(\frac{\delta + \hat{t}(v)}{\sqrt{\mathrm{Var}(\hat{t}(v))/n}} \right).$$

Intuitively, the p-value for the test is the maximum of all point-wise p-values because we are testing the maximum deviation of $\tilde{q}_1(v) - \tilde{q}_0(v)$.

Choosing an equivalence threshold δ So far, we have assumed that researchers have a clear idea what value should be used to assess the equivalence. When researcher have substantive knowledge about the appropriate value of δ given an application, it is reasonable to choose δ according to the knowledge. Oftentimes, this approach might not be feasible since it is not straightforward to form an idea of what value of δ should be deemed appropriate, especially when the value of δ is not directly tied to interpretable quantities such as causal effects. Although, any choice of δ is a valid choice because type I error is controlled for the corresponding null hypothesis, it seems useful to suggest a reasonable default value of δ to facilitate the practical use of the method.

I suggest the following value of δ as a reasonable starting point,

$$\delta_n = \min\left\{ 1.2\sqrt{\frac{n_1 + n_0}{n_1 n_0}},\ 1 \right\}$$

where $1.2 \approx \sqrt{-\log(\omega)/2}$ with $\omega = 0.05$ and $\sqrt{(n_1 + n_0)/(n_1 n_0)} \sim n^{-1/2}$ when $n_1 \sim n_0$. This is a threshold used in the conventional KS test which is a nonparametric test on the difference between two distribution functions. The test is based on the maximum difference between two cumulative distribution functions. In KS test, the value of ω by the level of a test, but it is fixed here. The key feature of this threshold is that δ_n depends on the sample size. The equivalence threshold that depends on the sample size is discussed in Romano (2005). Intuitively, this selection of δ implies that we raise the standard of what the equivalence means as the sample size increases. Therefore, rejecting the null with larger n will be a stronger evidence for the identification assumption.

4 Empirical Findings

In this section, I revisit the empirical application introduced in Section 2. We first reanalyze the two-wave panel of CCES (2010–2012). This two-wave panel is the data used for main analyses in the original studies. We then analyze the three-wave panel of CCES (2010–2012–2014) which allows us to assess the identification assumption using the pre-treatment periods.

In the following, we focus on estimating the following causal quantities:

$$\zeta_j = \mathbb{P}(Y_{i1}(1) = j \mid D_i = 1) - \mathbb{P}(Y_{i1}(0) = j \mid D_i = 1)$$

for $j = 0, 1, 2$. Recall that `less-strict` is coded as `0`, and `more-strict` category is coded as `2`.

4.1 Result from the two-wave panel

This section presents a result of the analysis on the two-wave sample from CCES ($n = 16620$). The outcome is measured in 2010 and 2012 and I treat a response in 2012 as the post-treatment outcome. Respondents living in a neighborhood where mass shootings happened within 100 miles between 2010 and 2012 are considered as treated ($n_1 = 4893$). In total, there were 16 mass shooting incidents recorded in the dataset between the two waves of CCES (Newman and Hartman, 2019, Appendix C). In addition to the analysis with the full sample, I also investigate effect heterogeneity by pre-treatment covariates. First, I investigate if the baseline safety of the neighborhood affects how people respond to mass shootings. Respondents are classified into either "prior exposure" group or "no prior exposure" group.

A respondent is in the “no prior exposure” group if she is living in a neighborhood that did not have mass shootings within 100 miles of the area for the last ten years (as of 2010). We would expect that people react differently to mass shootings depending on how frequent these events are in their life. Second, following the original papers, I investigate if effects vary across respondents’ party affiliations. Since issues related to gun control are debated along the party line in the US, we might expect that people react differently depending on which party they affiliate with.

Figure 3 shows the results. In the figure, circles represent point estimates for ζ_0 which can be interpreted as the causal effect on preferring less strict gun regulations, while triangles shows estimates for ζ_2 which captures the effect on preferring more strict control of firearm sales; squares are estimate for the middle category (ζ_1), which can be interpreted as a preference to the status quo. Along with point estimates, I also show the uncertain estimates.

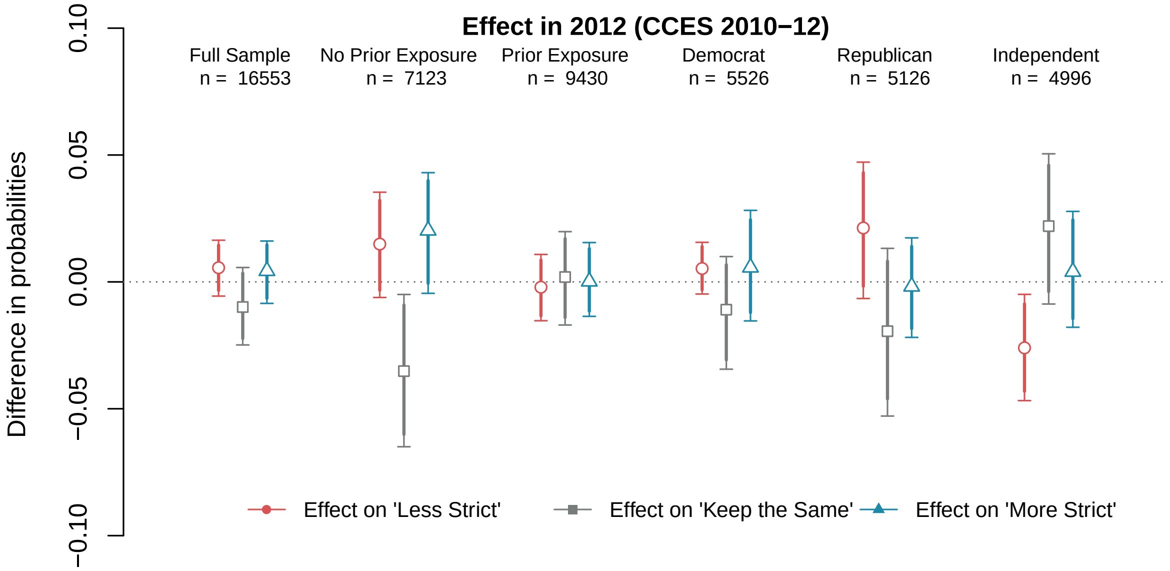

Figure 3: Estimated treatment effects with 90% (solid) and 95% (thick) confidence intervals. Circles indicate effect for `less-strict` (ζ_0), squares for `keep-the-same` (ζ_1) and triangles for `more-strict` (ζ_2). Labels above estimates indicate subsamples used for the analysis where n indicates the size of the sample.

Thick (thin) vertical lines show 90% (95%) confidence intervals calculated via block bootstraps. To account for the fact that the treatment assignment is at the zip code level, the bootstrap is conducted blocking at the zip code level. There are 9042 unique zip codes in the two-wave sample. I sample zip codes with replacement and create bootstrap samples. Confidence intervals are based on 2000 bootstrap iterations. Text labels shown above estimates indicate the subsamples and their sample sizes used for the analysis.

We can see that causal effect estimates are not statistically significant at the 10% level for all categories in the full sample. Estimates are precisely estimated and they are all close to zero, indicating that there is little evidence to suggest that mass shootings have, on average, any effect on the attitude towards gun control regulations among those who live in their vicinity of public shootings.

Following the original authors, I conduct two sets of subgroup analysis by "prior exposure" status and by the partisanship. The analysis reveals a similar pattern that most of the estimates are not statistically distinguishable from zero at the conventional level. However, we can also see that heterogeneity exists: the "no prior exposure" group has negative effect for the middle category ($\widehat{\zeta}_1 = -0.035$, SE $= 0.016$) which is statistically significant at the 5% level. This result implies that those living in the safer neighborhood (i.e., "no prior exposure") move away from the status quo. Although not statistically significant at the 10% level, we can also see that the effect on the less strict category is positively estimated for this "no prior exposure" group, indicating that the shift away from the status quo was probably not uni-directional. We can also see the negative effect on `less-strict` category among the independents ($\widehat{\zeta}_0 = -0.026$, SE $= 0.011$), which is statistically distinguishable from zero at the 5% level. In Appendix D, I present a result using a more granular measure of partisanship used in the survey, which asks respondents to categorize themselves on a 7-point scale from "strong Democrat" to "Strong Republican." The result in Figure D.1 shows that the effect is concentrated among lean Democrats, 91% of them categorized themselves as "independent" on the 3-point partisanship scale.

To further investigate the interactive effects between the the partisanship and the prior exposure status, I considered interactions between the two variables. Figure 4 shows the results of the analysis. As we can see the effect is concentrated among Democrats who are in the "no prior exposure" group, while none of the effects are statistically significant for other partisans. The figure also shows that partisanship does not play a role in the "prior exposure" group where estimates are indistinguishable from zero.

Finally, Barney and Schaffner (2019) consider different thresholds to determine who are "exposed" to the mass shootings. In addition to the above 100 mile threshold, I also estimate effects for the 25 mile threshold. The result is shown in Figure D.4 in Appendix D. We observe similar patterns with the previous results, while there are two notable differences.

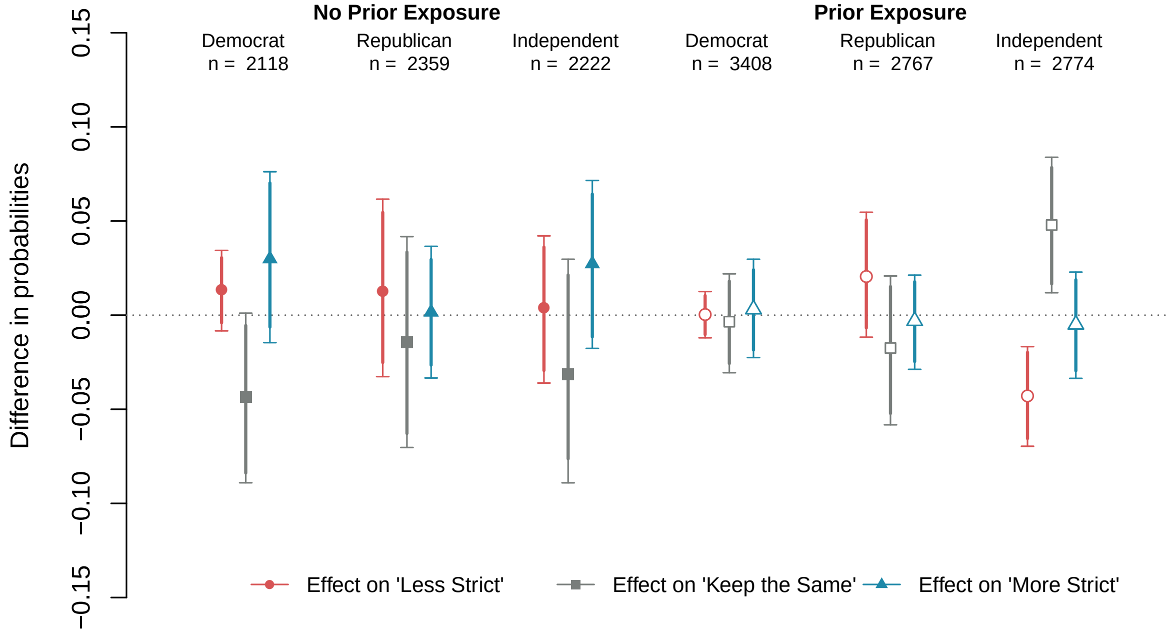

Figure 4: Estimated treatment effects with 90% (solid) and 95% (thick) confidence intervals. Circles indicate effects for `less-strict` (ζ_0), squares for `keep-as-they-are` (ζ_1) and triangles for `more-strict` (ζ_2). Text labels above the estimates indicate subsamples used for the analysis.

First, ζ_1 is now statistically significant at the 10% level ($\zeta_1 = -0.021$, SE $= 0.012$). Second, effects are clearer for Democrats without prior exposure: $\zeta_1 < 0$ and $\zeta_2 > 0$ and both of the estimates are statistically significant at the 5% level.

4.2 Diagnostics using three-wave panel

Next, I analyze the three-wave panel from CCES (2010-12-14) to assess if the identification assumption made in Assumption 3 is plausible or not. I subset the dataset so that I include only two types of respondents: those who experienced the mass shootings only after 2012 (treated group) and those who never experience the mass shootings throughout the sample periods (control group). This allows us to treat 2010 and 2012 as the pre-treatment periods, because no one in this subsample is affected by the treatment happened before 2012. To avoid the possibilities that the past exposure might affect the baseline attitudes, I further condition on the prior-exposure variable, including only respondents who are in the "no prior exposure" group. This subset consists of 2817 respondents among which 667 respondents are eventually treated between 2012 and 2014. In total, there were 28 incidents of mass shootings recorded in the dataset that happened between 2012 and 2014 waves.

I apply the diagnostic test proposed in Section 3.4 to the pre-treatment outcome. The goal here is to statistically test if the condition of the distributional parallel trend holds, namely, $\tilde{q}_1(v) = \tilde{q}_0(v)$ where $\tilde{q}_d(v)$ is the pre-treatment analog of the quantile-quantile relationship defined on group d (i.e., $q_d(v)$ in Assumption 3). Specifically, I test the null hypothesis of non-equivalence, $H_0 \colon \tilde{q}_1(v) \neq \tilde{q}_0(v)$ for all v against the equivalence.

I compute the test statistic $\hat{t}_{\max} = \max_v \hat{t}(v)$, where $\hat{t}(v) = \widehat{\tilde{q}}_1(v) - \widehat{\tilde{q}}_0(v)$, and corresponding confidence intervals at the 5% level. Each $\hat{t}(v)$ is computed by evaluating $\widehat{\tilde{q}}_1$ and $\widehat{\tilde{q}}_0$ on the finite number of grid points between 0.001 and 0.999 where the distance between points is set to 0.01. The equivalence threshold is chosen based on the heuristic criterion discussed in Section 3.4, $\delta_n = \sqrt{-\log(0.05)/2 \times n/(n_1 n_0)} \approx 0.054$ where $n = 2817$ and $n_1 = 667$.

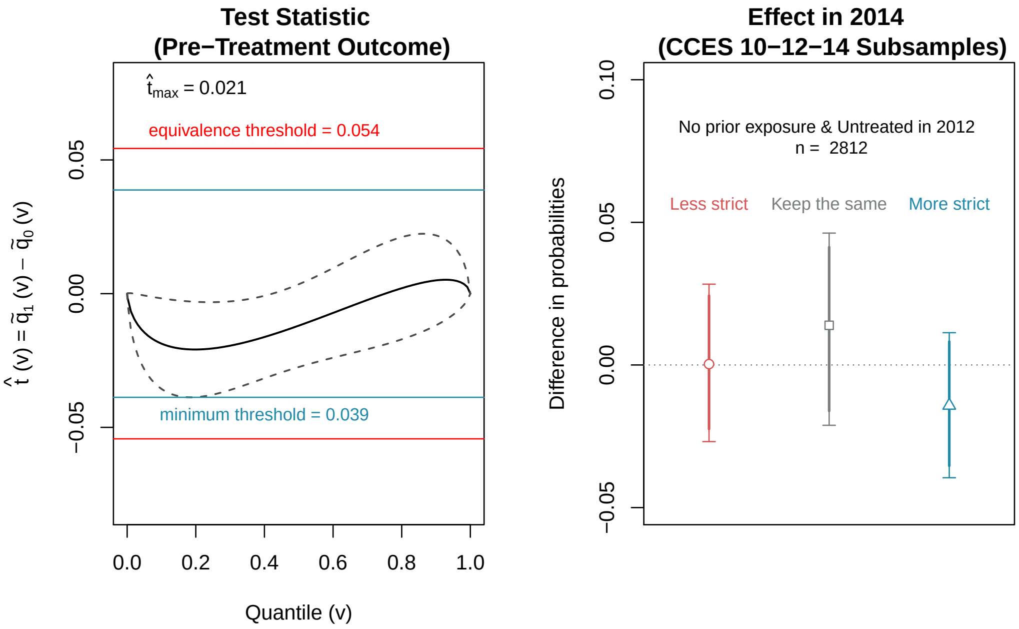

Figure 5: Left – Test statistics $\hat{t}(v)$ (solid line) with point-wise 95% confidence intervals (dashed lines). Red lines show the equivalence range $[-\delta_n, \delta_n]$. The figure shows that the largest (smallest) point of the upper (lower) confidence intervals is strictly contained in the equivalence range. It suggests that the null is rejected at the 5% level with $\delta_n = 0.054$. **Right** – Estimated causal effects with 90% (thick) and 95% (thin) confidence intervals. Either effects are not statistically distinguishable from zero at the 10% level.

The left panel of Figure 5 shows the difference between the two functions, $\hat{t}(v)$, evaluated at a value v on the unit interval (solid line). Dashed lines show the point-wise 95% confidence intervals. The test statistic (the estimated largest deviation) is $\hat{t}_{\max} = 0.021$ with the largest upper bound $\widehat{U}_{\max} = 0.022$ and the smallest lower bound $\widehat{L}_{\min} = -0.039$. Since the confidence range $[\widehat{L}_{\min}, \widehat{U}_{\max}]$ is strictly contained in the equivalence range $[-\delta_n, \delta_n]$, we can reject the null of non-equivalence at the 5% level ($p = 0.001$). In other words, for any choice of δ that is greater than $\max\{|\widehat{U}_{\max}|, |\widehat{L}_{\min}|\} = 0.039$, we reject the null at the 5% level. The result suggests that during the pre-treatment periods the data supports the analogous condition of Assumption 3.

After confirming the plausibility of the key identification assumption, we now analyze the outcome measured in 2012 (pre-treatment) and 2014 (post-treatment) to estimate the causal effect for the three-wave subsample. The right panel of Figure 5 shows the result of the analysis. We can see that none of the estimates are statistically distinguishable from zero at the 10% level ($\widehat{\zeta}_0 = 0.000$, SE $= 0.0144$; $\widehat{\zeta}_1 = 0.0140$, SE $= 0.0172$; and $\widehat{\zeta}_2 = -0.0142$, SE $= 0.0132$). This result somewhat contradicts with findings in the previous section, where I found a negative effect on ζ_1 among the "no prior exposure" group. Although there are many possible reasons why effects could vary over time. One possibility is simply the size of the dataset. The subset of the three-wave panel has smaller respondents than the two-wave samples analyzed in the previous section. This difference obviously translates into differences in uncertainty estimates. Another possibility is due to the contextual differences. On December 14th, after the wave of CCES 2012, the Sandy Hook Elementary School shooting happened. This was one of the deadliest mass shootings in the US history, which possibly raised the salience of the issue affecting gun control regulations in many states.

5 Concluding Remarks

In spite of the recent developments in the literature on the DID design, less attention has been paid when the outcome is measured on an ordinal scale. In this paper, I proposed a method that allows scholars to leverage ordinal outcomes without making the linearity assumption as in the the standard DID analysis. I also proposed a procedure that assesses if the key identification assumption is plausible when additional pre-treatment periods are available. This enable scholars to inspect the data and to discuss if the assumption is reasonable given a particular dataset they analyze, which is a crucial step for any research that attempts to establish a causal relationship.

Several extensions of the proposed methods are possible. In Appendix B, I demonstrate that the proposed method can be useful to estimate other types of causal estimands such as the proportion of who benefits from the treatment (e.g. Lu et al., 2018). Recent years, it has been argued that such estimands are preferable because the distributional treatment effects considered in the main text are not necessarily easy to interpret. Since the proposed method identifies the entire distributions of the potential outcomes, it is possible to compute any causal estimand that is a function of marginal distributions of the potential outcome. In the appendix, I also discuss how to incorporate time-varying covariates, which requires a further modeling assumption.

In Appendix F, I present a simulation study where I assess the finite sample property of the proposed estimator and the testing procedure. I find that under the correct model specification, the proposed method outperforms the two competing methods: the standard difference-in-differences on the dichotomized outcome and the ordered probit model. I also find that the type-I error is properly controlled for the proposed testing procedure, while the power depends on the choice of the equivalence threshold.

Finally, future research should consider an extension to a complex design. Specifically, the staggered adoption design where the treatment is assigned overtime is a popular data structure in applied works. Although the development of such methods is beyond the scope of this paper, those methodologies should improve the analysis of ordinal outcome beyond the standard DID setting.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2018). Synthetic difference in differences. *arXiv preprint arXiv:1812.09970*.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Barney, D. J. and Schaffner, B. F. (2019). Reexamining the effect of mass shootings on public support for gun control. *British Journal of Political Science*, 49(4):1555–1565.

Callaway, B., Li, T., and Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413.

Callaway, B. and Sant'Anna, P. H. (2018). Difference-in-differences with multiple time periods and an application on the minimum wage and employment. *arXiv preprint arXiv:1803.09015*.

Card, D. and Krueger, A. (1994). Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4):772–793.

Chiba, Y. (2017). Sharp nonparametric bounds and randomization inference for treatment effects on an ordinal outcome. *Statistics in Medicine*, 36(25):3966–3975.

Egami, N. and Yamauchi, S. (2019). How to improve the difference-in-differences design with multiple pre-treatment periods. *Working Paper*.

Glynn, A. and Ichino, N. (2019). Generalized nonlinear difference-in-difference-in-differences. *Working Paper*.

Hartman, T. K. and Newman, B. J. (2019). Accounting for pre-treatment exposure in panel data: Re-estimating the effect of mass public shootings. *British Journal of Political Science*, 49(4):1567–1576.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* John Wiley & Sons.

Lechner, M. et al. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224.

Lee, M.-j. (2016). Generalized difference in differences with panel data and least squares estimator. *Sociological Methods & Research*, 45(1):134–157.

Li, F. (2019). Double-robust estimation in difference-in-differences with an application to traffic safety evaluation. *arXiv preprint arXiv:1901.02152.*

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, pages 44–53.

Liu, W., Bretz, F., Hayter, A., and Wynn, H. (2009). Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region. *Biometrics*, 65(4):1279–1287.

Lu, C., Nie, X., and Wager, S. (2019). Robust nonparametric difference-in-differences estimation. *arXiv preprint arXiv:1905.11622.*

Lu, J. (2018). On the partial identification of a new causal measure for ordinal outcomes. *Statistics & Probability Letters*, 137:1–7.

Lu, J., Ding, P., and Dasgupta, T. (2018). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43(5):540–567.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Newman, B. J. and Hartman, T. K. (2019). Mass shootings and public support for gun control. *British Journal of Political Science*, 49(4):1527–1553.

Qin, J. and Zhang, B. (2008). Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):329–349.

Romano, J. P. (2005). Optimal testing of equivalence hypotheses. *The Annals of Statistics*, 33(3):1036–1047.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen, E. J. T. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science*, 31(3):348.

Volfovsky, A., Airoldi, E. M., and Rubin, D. B. (2015). Causal inference for ordinal outcomes. *arXiv preprint arXiv:1501.01234*.

Appendix

A Proofs of Propositions

A.1 Lemmas

Before proving propositions, we present useful lemmas.

Lemma 1 (Identification of mean and variance of the latent variables.)**.** Suppose that the cutoffs are fixed at κ_1 and κ_2 for $Y_{dt} = j \in \{0, 1, 2\}$. Then, μ_{dt} and σ_{dt} in $Y^*_{dt} \sim \mu_{dt} + \sigma_{dt} U$ are uniquely identified from the observed probability distribution.

Proof of Lemma 1. Suppose that U has the density $f_U(u)$. Then, we can form a non-linear system of equations

$$\begin{aligned}
\Pr(Y_{dt} = 0) &= \int_{-\infty}^{\kappa_1} f_U((y^* - \mu_{dt})/\sigma_{dt}) dy^* \\
\Pr(Y_{dt} = 2) &= \int_{\kappa_2}^{\infty} f_U((y^* - \mu_{dt})/\sigma_{dt}) dy^*
\end{aligned}$$

which are sufficient for estimating μ and σ. □

Lemma 2 (Alternative formula for identification)**.** Suppose $Y_{dt} = j \in \{0, 1, 2\}$. Let $v_1 = F_{01}(\kappa_1)$ and $v_2 = F_{01}(\kappa_2)$ where $\boldsymbol{\kappa}$ is a set of fixed cutoffs. Under Assumption 1, 2 and 3, we identify μ_{11} and σ_{11} by the following system of non-linear equations:

$$\begin{aligned}
q_0(v_1) &= \int_{-\infty}^{F_{10}^{-1}(v_1)} f_U((y^* - \mu_{11})/\sigma_{11}) dy^* \\
q_0(v_2) &= \int_{-\infty}^{F_{10}^{-1}(v_2)} f_U((y^* - \mu_{11})/\sigma_{11}) dy^*.
\end{aligned}$$

Proof of Lemma 2. Under the distributional parallel trends assumption, we have $q_0(v) = q_1(v)$ for all $v \in [0, 1]$. Then,

$$\begin{aligned}
q_0(v) &= F_{11} \circ F_{10}^{-1}(v) \\
&= \int_{-\infty}^{F_{10}^{-1}(v)} f_{11}(y^*) dy^* \\
&= \int_{-\infty}^{F_{10}^{-1}(v)} f_U((y^* - \mu_{11})/\sigma_{11}) dy^*
\end{aligned}$$

where the first equality is due to the definition of $q_d(v)$ and the last equality follows by Assumption 2. Pick v_1 and v_2 as in the statement. Drawing on a similar to the argument in Lemma 1, we obtain the identification.

□

Lemma 3 (Invariance of $\widehat{\zeta}_j$ under different cutoffs)**.** Suppose we have two sets of cutoffs $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}'$ ($\boldsymbol{\kappa} \neq \boldsymbol{\kappa}'$) for $Y_{dt} = j \in \{0, 1, 2\}$. Then, $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\kappa}) = \widehat{\boldsymbol{\zeta}}(\boldsymbol{\kappa}')$.

Proof of Lemma 3. Let $F_{dt}(y) = \Pr(Y^*_{dt} \leq y)$ denote the cumulative distribution function of the latent variable Y^*_{dt} under the cutoff $\boldsymbol{\kappa}$. Similarly, let $\widetilde{F}_{dt}(y)$ denote the CDF under $\boldsymbol{\kappa}'$. To show the causal effect estimates are invariant to the choice of cutoffs, it is sufficient to show that $F_{11}(y) = \widetilde{F}_{11}(y)$, that is, the invariance of the identified counterfactual latent distribution.

We first show that $F_{00}(F^{-1}_{01}(u)) = \widetilde{F}_{00}(\widetilde{F}^{-1}_{01}(u'))$ for $u = F_{01}(\kappa_1)$ and $u' = \widetilde{F}_{01}(\kappa'_1)$. Now, note that we have $u = u'$ because

$$\begin{aligned} F_{01}(\kappa_1) &= \Pr(Y_{01} = 0) \\ &= \widetilde{F}_{01}(\kappa'_1) \end{aligned}$$

where Y_{01} is the observed outcome. Thus,

$$\begin{aligned} F_{00}(F^{-1}_{01}(u)) - u &= F_{00}(\kappa_1) - F_{01}(\kappa_1) \\ &= \Pr(Y_{00} = 0) - \Pr(Y_{01} = 0) \\ &= \widetilde{F}_{00}(\kappa_1) - \widetilde{F}_{01}(\kappa_1) \\ &= \widetilde{F}_{00}(\widetilde{F}^{-1}_{01}(u')) - u' \end{aligned}$$

which proves that $F_{00}(F^{-1}_{01}(u)) = \widetilde{F}_{00}(\widetilde{F}^{-1}_{01}(u'))$.

Next, by the similar argument, we have that $F_{10}(\kappa_1) = \widetilde{F}_{10}(\kappa'_1)$, because

$$\begin{aligned} F_{10}(\kappa_1) &= \Pr(Y_{10} = 0) \\ &= \widetilde{F}_{10}(\kappa'_1). \end{aligned}$$

Repeating the above two steps for κ_2 and κ'_2, we obtain that

$$\begin{aligned} F_{00}(F^{-1}_{01}(u)) &= \int_{-\infty}^{F^{-1}_{10}(u)} f_U((y^* - \mu_{11})/\sigma_{11}) dy^* \\ &= \widetilde{F}_{00}(\widetilde{F}^{-1}_{01}(u')) \end{aligned}$$

and

$$\begin{aligned} F_{00}(F^{-1}_{01}(v)) &= \int_{-\infty}^{F^{-1}_{10}(v)} f_U((y^* - \mu_{11})/\sigma_{11}) dy^* \\ &= \widetilde{F}_{00}(\widetilde{F}^{-1}_{01}(v')) \end{aligned}$$

where $v = F_{01}(\kappa_2)$ and $v' = \widetilde{F}_{01}(\kappa'_2)$.

Applying the result of Lemma 2, we can see that μ_{11} and σ_{11} are uniquely identified under different sets of cutoffs, that is, $\boldsymbol{\zeta}(\boldsymbol{\kappa}) = \boldsymbol{\zeta}(\boldsymbol{\kappa}')$.

Finally, replacing all quantities with their sample analog, we conclude that $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\kappa}) = \widehat{\boldsymbol{\zeta}}(\boldsymbol{\kappa}')$.

□

Lemma 4 (Asymptotic Normality of Causal Estimates)**.** Under some regularity conditions, as $n \to \infty$ with $n_1/n \to k$, we have that

$$\sqrt{n}(\widehat{\zeta}_j - \zeta_j) \rightsquigarrow \mathcal{N}(0, \sigma_j^2) \tag{A.1}$$

Proof of Lemma 4. We prove the statement by showing the following two statements:

1. $\sum_{i=1}^{n} D_i \mathbf{1}\{Y_{i1} = j\}/n_1$ is $\sqrt{n}$-consistent estimator for $\Pr(Y_{i1}(1) = j \mid D_i = 1)$.
2. $\widehat{\boldsymbol{\theta}}_{11} = (\widehat{\mu}_{11}, \widehat{\sigma}_{11})^\top$ is $\sqrt{n}$-consistent estimator for $\boldsymbol{\theta}_{11}$.

We then use the continuous mapping theorem for the convergence in distribution to obtain the final result.

To be clear, we (sometime implicitly) condition on $D_i = 1$ throughout the proof, which assumes that there is a super population of units with $D_i = 1$. Now let $W_i = D_i \mathbf{1}\{Y_{i1} = j\}$ and $\pi_{11}(j) = \Pr(Y_{i1}(1) = j \mid D_i = 1)$. Under the assumption that $Y_{it} \perp\!\!\!\perp Y_{i't'}$ for any combination of i and t, it follows that $\sum_{i=1}^{n} W_i/n_1 \to \mathbb{E}[\mathbf{1}\{Y_{i1}(1) = j\} \mid D_i = 1] = \pi_{11}(j)$ as $n \to \infty$ with $n_1/n \to k$ by the law of large numbers. This proves the consistency. By the central limit theorem, it also follows that

$$\begin{aligned}
\sqrt{n}\left(\frac{1}{n_1}\sum_{i=1}^{n} W_i - \pi_{11}(j)\right) &= \frac{n}{n_1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i - \pi_{11}(j)) \\
&\rightsquigarrow \mathcal{N}\left(0, \frac{\mathrm{Var}(\widehat{\pi}_{11}(j))}{k^2}\right)
\end{aligned}$$

as $n \to \infty$ with $n_1/n \to k$.

Next, recall that $\widehat{\boldsymbol{\theta}}_{11}$ is given as a transformation of $\widehat{\boldsymbol{\theta}}_{00}$, $\widehat{\boldsymbol{\theta}}_{01}$ and $\widehat{\boldsymbol{\theta}}_{10}$, all of which are MLE of the problem described in Section 3.3. Therefore, under the assumption of correct model specification, we obtain that $\widehat{\boldsymbol{\theta}}_{00}$, $\widehat{\boldsymbol{\theta}}_{01}$ and $\widehat{\boldsymbol{\theta}}_{10}$ are jontly asymptotically normal centered around $\boldsymbol{\theta}_{00}$, $\boldsymbol{\theta}_{01}$ and $\boldsymbol{\theta}_{10}$. By the continuous mapping theorem, it follows that $\widehat{\boldsymbol{\theta}}_{11}$ is also asymptotically normally distributed.

Finally, using the continuous mapping theorem again, we have that $\widehat{\zeta}_j$ is a $\sqrt{n}$ consistent estimator for ζ_j where the variance σ_j is obtained by the delta method (See Lemma 6).

□

Lemma 5 (Asymptotic Normality of $\boldsymbol{\theta}$ for Pre-treatment Parameters)**.** Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_{00}^\top, \boldsymbol{\theta}_{01}^\top, \boldsymbol{\theta}_{10}^\top, \boldsymbol{\theta}_{11}^\top)^\top$, all of which are estimated using the data from the pre-treatment periods. Then, under regularity conditions in Newey and McFadden (1994), the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with covariance Ω,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightsquigarrow \mathcal{N}(0, \Omega) \tag{A.2}$$

where Ω is a block-diagonal matrix under independence assumption.

Proof of Lemma 5. The result is a direct application of the standard result of the maximum likelihood estimation. Therefore, proof is omitted. □

Lemma 6 (Asymptotic Distribution of the Test Statistic)**.** Assume that $U \sim \mathcal{N}(0,1)$. Let $t(v;\boldsymbol{\theta}) = \tilde{q}_1(v;\boldsymbol{\theta}) - \tilde{q}_0(v;\boldsymbol{\theta})$ and $\hat{t}(v) \equiv t(v;\widehat{\boldsymbol{\theta}})$. Then, we have that

$$\sqrt{n}(t(v;\widehat{\boldsymbol{\theta}}) - t(v;\boldsymbol{\theta})) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}(\hat{t}(v))) \tag{A.3}$$

for each $v \in [0,1]$ with

$$\mathrm{Var}(\hat{t}(v)) = \left(\frac{\partial}{\partial \boldsymbol{\theta}} t(v;\boldsymbol{\theta})\right)^{\top} \Omega \left(\frac{\partial}{\partial \boldsymbol{\theta}} t(v;\boldsymbol{\theta})\right) \tag{A.4}$$

where $\boldsymbol{\theta}$ is evaluated at the truth, Ω is the asymptotic variance covariance matrix of $\widehat{\boldsymbol{\theta}}$ given in Lemma 5, and the gradient takes the form of

$$\frac{\partial}{\partial \boldsymbol{\theta}} t(v;\boldsymbol{\theta}) = \begin{bmatrix} \exp(-z_0^2)/\sqrt{2\pi}\sigma_{00} \\ \exp(-z_0^2)z_0/\sqrt{\pi}\sigma_{00} \\ -\exp(-z_0^2)/\sqrt{2\pi}\sigma_{00} \\ -\exp(-z_0^2)\mathrm{erf}^{-1}(2v-1)/\sqrt{\pi}\sigma_{00} \\ -\exp(-z_1^2)/\sqrt{2\pi}\sigma_{10} \\ -\exp(-z_1^2)z_1/\sqrt{\pi}\sigma_{10} \\ \exp(-z_1^2)/\sqrt{2\pi}\sigma_{10} \\ \exp(-z_1^2)\mathrm{erf}^{-1}(2v-1)/\sqrt{\pi}\sigma_{10} \end{bmatrix}$$

with

$$z_d \equiv \frac{\mu_{d1} - \mu_{d0}}{\sigma_{d0}\sqrt{2}} + \frac{\mathrm{erf}^{-1}(2v-1)}{\sigma_{d0}/\sigma_{d1}}.$$

Proof of Lemma 6. Recall that the derivative of the error function is given by

$$\frac{d}{dz}\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} e^{-z^2} \tag{A.5}$$

which is differentiable with respect to z.

Now, I compute the derivative of $t(v;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$,

$$\frac{\partial}{\partial\boldsymbol{\theta}}t(v;\boldsymbol{\theta}) = \begin{bmatrix} \left(\frac{\partial}{\partial\boldsymbol{\theta}_{00}}t(v;\boldsymbol{\theta})\right) \\ \left(\frac{\partial}{\partial\boldsymbol{\theta}_{01}}t(v;\boldsymbol{\theta})\right) \\ \left(\frac{\partial}{\partial\boldsymbol{\theta}_{10}}t(v;\boldsymbol{\theta})\right) \\ \left(\frac{\partial}{\partial\boldsymbol{\theta}_{11}}t(v;\boldsymbol{\theta})\right) \end{bmatrix} = \begin{bmatrix} \exp(-z_0^2)/\sqrt{2\pi}\sigma_{00} \\ \exp(-z_0^2)z_0/\sqrt{\pi}\sigma_{00} \\ -\exp(-z_0^2)/\sqrt{2\pi}\sigma_{00} \\ -\exp(-z_0^2)\mathrm{erf}^{-1}(2v-1)/\sqrt{\pi}\sigma_{00} \\ -\exp(-z_1^2)/\sqrt{2\pi}\sigma_{10} \\ -\exp(-z_1^2)z_1/\sqrt{\pi}\sigma_{10} \\ \exp(-z_1^2)/\sqrt{2\pi}\sigma_{10} \\ \exp(-z_1^2)\mathrm{erf}^{-1}(2v-1)/\sqrt{\pi}\sigma_{10} \end{bmatrix}$$

where

$$z_d \equiv \frac{\mu_{d1}-\mu_{d0}}{\sigma_{d0}\sqrt{2}} + \frac{\mathrm{erf}^{-1}(2v-1)}{\sigma_{d0}/\sigma_{d1}}$$

Given that Lemma 5 establishes the asymptotic normality of $\widehat{\boldsymbol{\theta}}$, the result immediately follows by the application of the Delta method. Then, we get

$$\sqrt{n}(t(v;\widehat{\boldsymbol{\theta}}) - t(v;\boldsymbol{\theta})) \rightsquigarrow \left(\frac{\partial}{\partial\boldsymbol{\theta}}t(v;\boldsymbol{\theta})\right)\mathcal{N}(0,\Omega) \tag{A.6}$$

From here, we obtain the variance formula as

$$\begin{aligned} \mathrm{Var}(\hat{t}(v)) = &\left(\frac{\partial}{\partial\boldsymbol{\theta}_{00}}t(v)\right)^{\top}\Omega_{00}\left(\frac{\partial}{\partial\boldsymbol{\theta}_{00}}t(v)\right) + \left(\frac{\partial}{\partial\boldsymbol{\theta}_{01}}t(v)\right)^{\top}\Omega_{01}\left(\frac{\partial}{\partial\boldsymbol{\theta}_{01}}t(v)\right) \\ &+ \left(\frac{\partial}{\partial\boldsymbol{\theta}_{10}}t(v)\right)^{\top}\Omega_{10}\left(\frac{\partial}{\partial\boldsymbol{\theta}_{10}}t(v)\right) + \left(\frac{\partial}{\partial\boldsymbol{\theta}_{11}}t(v)\right)^{\top}\Omega_{11}\left(\frac{\partial}{\partial\boldsymbol{\theta}_{11}}t(v)\right) \end{aligned}$$

where $\hat{t}(v) \equiv t(v;\widehat{\theta})$. □

Lemma 7 (Validity of $(1-\alpha)$ level sets (Liu et al., 2009))**.** Let $t(v) = \tilde{q}_2(v) - \tilde{q}_1(v)$. Suppose $\widehat{U}_{1-\alpha}(v)$ and $\widehat{L}_{1-\alpha}(v)$ are point-wise upper and lower $(1-\alpha)$ level confidence intervals, respectively such that $\widehat{U}_{1-\alpha}(v) = \hat{t}(v) + \Phi^{-1}(1-\alpha)\sqrt{\mathrm{Var}(\hat{t}(v))/n}$ and $\widehat{L}_{1-\alpha}(v) = \hat{t}(v) - \Phi^{-1}(1-\alpha)\sqrt{\mathrm{Var}(\hat{t}(v))/n}$. Then,

$$\mathbb{P}\left(\max_{v\in[0,1]} t(v) \leq \max_{v'\in[0,1]} \widehat{U}_{1-\alpha}(v')\right) \geq 1-\alpha \tag{A.7}$$

$$\mathbb{P}\left(\min_{v\in[0,1]} t(v) \geq \min_{v'\in[0,1]} \widehat{L}_{1-\alpha}(v')\right) \geq 1-\alpha \tag{A.8}$$

Proof of Lemma 7. Recall that $\widehat{U}_{1-\alpha}(v)$ is a point-wise $100(1-\alpha)\%$ level confidence interval.

This implies that

$$\mathbb{P}(t(v) \leq \widehat{U}_{1-\alpha}(v)) = 1 - \alpha$$

for any $v \in [0,1]$. Now, let $v^* = \arg\max_{v} t(v)$. Then, we have that

$$1 - \alpha = \mathbb{P}(t(v^*) \leq \widehat{U}_{1-\alpha}(v^*)) \leq \mathbb{P}(t(v^*) \leq \max_{v'} \widehat{U}_{1-\alpha}(v'))$$

which proves that $\mathbb{P}(\max_v t(v) \leq \max_{v'} \widehat{U}_{1-\alpha}(v')) \geq 1 - \alpha$.

□

A.2 Proofs

Proof of Proposition 1. Let U denote a random variable with mean 0 and variance 1 and denote its cumulative distribution function by F_U. For $v \sim \mathcal{U}(0,1)$, we have

$$\begin{aligned} q_0(v) &\equiv F_{Y^*_{00}} \circ F^{-1}_{Y^*_{01}}(v) \\ &= F_U\left(\frac{\mu_{01} - \mu_{00}}{\sigma_{00}} + \frac{\sigma_{01}}{\sigma_{00}} F_U^{-1}(v)\right) \end{aligned}$$

The equality in the above expression holds because Y^*_{dt} follows the location-scale family, which implies

$$\begin{aligned} F_{Y^*_{dt}}(y^*) &= F_U\left(\frac{y^* - \mu_{dt}}{\sigma_{dt}}\right) \\ F^{-1}_{Y^*_{dt}}(v) &= \mu_{dt} + \sigma_{dt} F_U^{-1}(v) \end{aligned}$$

By Assumption 3,

$$\begin{aligned} F^{-1}_{Y^*_{11}}(v) &= F^{-1}_{Y^*_{10}}(q_0(v)) \\ &= \mu_{10} + \sigma_{10} F_U^{-1}(q_0(v)) \\ &= \mu_{10} + \sigma_{10}\left(\frac{\mu_{01} - \mu_{00}}{\sigma_{00}} + \frac{\sigma_{01}}{\sigma_{00}} F_U^{-1}(v)\right) \\ &\equiv \mu_{11} + \sigma_{11} F_U^{-1}(v) \end{aligned}$$

where

$$\mu_{11} \equiv \mu_{10} + \frac{\mu_{01} - \mu_{00}}{\sigma_{00}/\sigma_{10}} \quad \text{and} \quad \sigma_{11} \equiv \frac{\sigma_{10}\sigma_{01}}{\sigma_{00}}.$$

Combined with the fact that U is a continuous and parametric distribution, we recovers the distribution of Y^*_{11}.

□

Proof of Proposition 2. Case 1: $t \geq \epsilon$. In this case, the test makes a "mistake" because the upper bound does not cover ϵ. Thus, we can focus on an event $\{\max_v \widehat{U}_{1-\alpha}(v) < \epsilon\}$. Since $\mathbb{P}(\max_v \widehat{U}_{1-\alpha}(v) < \epsilon) = \mathbb{P}(\widehat{U}_{1-\alpha}(v) < \epsilon, \forall v)$, we can bound $\mathbb{P}(\max_v \widehat{U}_{1-\alpha}(v) < \epsilon)$ as

$$\mathbb{P}(\max_v \widehat{U}_{1-\alpha}(v) < \epsilon) \leq \min_v \mathbb{P}(\widehat{U}_{1-\alpha}(v) < \epsilon) \tag{A.9}$$

Now, consider a particular value of v. Then, we have that

$$\begin{aligned}\mathbb{P}(\widehat{U}_{1-\alpha}(v) < \epsilon) &\leq \mathbb{P}(\widehat{U}_{1-\alpha}(v) < t) \\ &= 1 - \mathbb{P}(t \leq \widehat{U}_{1-\alpha}(v)) \\ &\leq 1 - (1-\alpha) = \alpha \quad (n \to \infty)\end{aligned}$$

where the last inequality uses the fact that asymptotically $\widehat{U}_{1-\alpha}$ is a $(1-\alpha)$ level confidence interval (Lemma 7).

Case 2: $t \leq -\epsilon$. In this case, we focus on the other event $\{\min_v \widehat{L}_{1-\alpha}(v) \geq -\epsilon\}$. Since $\mathbb{P}(\inf_v \widehat{L}_{1-\alpha}(v) \geq -\epsilon) \leq \min_v \mathbb{P}(\widehat{L}_{1-\alpha}(v) \geq -\epsilon)$, we have that

$$\begin{aligned}\mathbb{P}(\widehat{L}_{1-\alpha}(v) \geq -\epsilon) &\leq \mathbb{P}(\widehat{L}_{1-\alpha}(v) \geq t) \\ &= 1 - \mathbb{P}(\widehat{L}_{1-\alpha}(v) \leq t) \\ &\leq 1 - (1-\alpha) = \alpha \quad (n \to \infty)\end{aligned}$$

where the last inequality is a direct application of Lemma 7. □

B Extensions

B.1 Other estimands

This section provides an extension of the proposed methodology. Following the recent developments in the literature on causal inference with ordinal outcome, where more interpretable estimands have been proposed, I show how to apply the proposed method to those new estimands.

In the above section, we have focused on a particular causal estimand, Δ_j which is a difference in probabilities defined for a specific reference category j. One issue of this quantity Δ_j is that depending on the choice of reference category j, the sign of the estimate might flip. This means that interpretation becomes tricky because it is completely possible to observe $\Delta_j > 0$ and $\Delta_{j'} < 0$ for $j \neq j'$ with the same data. From this, we cannot even conclude that the treatment had "positive" effect or not.

To circumvent this problem associated with Δ_j, recent papers turn to different kinds of estimands for ordinal outcome (e.g., Volfovsky et al., 2015; Chiba, 2017; Lu et al., 2018; Lu, 2018). For example, Lu et al. (2018) considers the following estimand:

$$\eta = \mathbb{P}(Y_{i1}(1) \geq Y_{i1}(0) \mid D_i = 1) \tag{B.1}$$

This is a proportion units of who benefit from (or at least not harmed by) the treatment. In our example, η captures the proportion of treated respondents who change their opinion toward gun control (regardless of their baseline attitudes) after experiencing the mass shooting in their neighborhood. This quantity is easy to interpret because it does not depend on the baseline attitude and smaller value of η indicates that there are few respondents who change their opinion towards gun controls. However, since η depends on the joint distribution of potential outcomes, $(Y_{i1}(0), Y_{i1}(1))$, it cannot be point identified. Lu et al. (2018) provides a closed form bound for this estimand using only the marginal distribution of $Y_{i1}(1)$ and $Y_{i1}(0)$.

The benefit of the proposed method over the dichotomizing-the-outcome approach is that it identifies the entire distribution of $Y_{i1}(0) \mid D_i = 1$, whereas the information about the entire distribution is lost when we use the coarsened outcome. This implies that we can estimate the bound based on $\widehat{\theta}_{11}$ given in Equation (3.8). Following the result of Lu et al. (2018), we can estimate the bound $[\widehat{\underline{\eta}}, \widehat{\overline{\eta}}]$ as

$$\widehat{\underline{\eta}} = \max_{0 \leq j \leq J-1} \left\{ [\Phi(\kappa_{j+1} \mid \widehat{\theta}_{11}) - \Phi(\kappa_j \mid \widehat{\theta}_{11})] + \widehat{\Delta}_j \right\} \quad \text{and} \quad \widehat{\overline{\eta}} = 1 + \min_{0 \leq j \leq J-1} \widehat{\Delta}_j$$

where $\widehat{\Delta}_j$ is the estimate of the distributional effect, and $\widehat{\Delta}_0 = 0$ by construction. We can see from the formula that the upper bound is informative as long as there is at least one reference category j such that $\widehat{\Delta}_j < 0$ for $j = 1, \ldots, J-1$. Otherwise, $\Delta_0 = 0$ will be the minimum and thus we get the non-informative upper bound $\widehat{\overline{\eta}} = 1$.

B.2 Time-varying covariates

Researchers might want to incorporate time-varying covariates into the analysis to further gain efficiency. I discuss that the parametric specification of the proposed method allows the use of such covariates for analysis. Although sometimes appealing, I emphasize that parametric specification introduces additional assumptions for the analysis.

Let $\mathbf{X}_{it} \in \mathbb{R}^p$ denote a p dimensional vector of time varying covariates. We can model the mean and the variance of the latent utilities as

$$\mu_{it} = \mathbf{Z}_{it}^\top \boldsymbol{\gamma}_0 \quad \text{and} \quad \sigma_{it} = \exp(\mathbf{Z}_{it}^\top \boldsymbol{\gamma}_1)$$

where $\mathbf{Z}_{it} = (1, D_i, t, D_i \cdot t, \mathbf{X}_{it}^\top)^\top$. Then, we can express the observed choice probability as

$$\mathbb{P}(Y_{it} = j \mid \mathbf{Z}_{it}) = \Phi(\kappa_{j+1} \mid \mu_{it}, \sigma_{it}) - \Phi(\kappa_j \mid \mu_{it}, \sigma_{it}).$$

We estimate parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}_1^\top)^\top$ by the maximum likelihood.

$$\widehat{\gamma} = \arg\max_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \sum_{t=0}^{1} \sum_{j=0}^{J-1} \mathbf{1}\{Y_{it} = j\} \log \Big\{ \Phi(\kappa_{j+1} \mid \mathbf{Z}_{it}, \boldsymbol{\gamma}) - \Phi(\kappa_j \mid \mathbf{Z}_{it}, \boldsymbol{\gamma}) \Big\}.$$

Finally, the quantities of interest is estimated by taking the sample average of predicted probabilities.

$$\widehat{\Delta}_j = \frac{1}{n_1} \sum_{i=1}^{n} D_i \Big\{ \mathbb{P}(Y_{i1} \geq j \mid D_i = 1, \mathbf{X}_{i1}, \widehat{\gamma}) - \mathbb{P}(Y_{i1} \geq j \mid D_i = 0, \mathbf{X}_{i1}, \widehat{\gamma}) \Big\}$$

Note that the marginalization of covariates is with respect to the distribution for the treated, because our estimand Δ_j is defined for the treated units.

C Dichotomizing the Outcome: An Example

Coarsening the ordinal outcome into a binary variable is a common practice often employed in applied works. Although this procedure allows scholars to utilize the standard linear DID, I will demonstrate in this section that this operation leads to an inconsistent result depending on how the new variable is created.

To see this, let's consider a simple example with three categories, $Y_{it} \in \{0, 1, 2\}$. There are two possible ways to transform this variable into a binary outcome, $\widetilde{Y}_{it} = \mathbf{1}\{Y_{it} = 2\}$ or $\check{Y}_{it} = \mathbf{1}\{Y_{it} \geq 1\}$. Under this setup, we require two separate parallel trends assumptions for identification,

$$\begin{aligned}
&\text{PT1: } \mathbb{E}[\widetilde{Y}_{i1}(0) - \widetilde{Y}_{i0}(0) \mid D_i = 1] = \mathbb{E}[\widetilde{Y}_{i1}(0) - \widetilde{Y}_{i0}(0) \mid D_i = 0] \\
&\text{PT2: } \mathbb{E}[\check{Y}_{i1}(0) - \check{Y}_{i0}(0) \mid D_i = 1] = \mathbb{E}[\check{Y}_{i1}(0) - \check{Y}_{i0}(0) \mid D_i = 0]
\end{aligned}$$

PT1 identifies $\Delta_2 = \Pr(Y_{i1}(1) = 2 \mid D_i = 1) - \Pr(Y_{i1}(0) = 2 \mid D_i = 1)$ and PT2 identifies $\Delta_1 = \Pr(Y_{i1}(1) \geq 1 \mid D_i = 1) - \Pr(Y_{i1}(0) \geq 1 \mid D_i = 1)$. Also let $\pi_{j|d}^{(t)} = \Pr(Y_{it}(0) = j \mid D_i = d)$ be the conditional probability for the potential outcome under the control.

Now consider the following data generating process which specifies the marginal distributions for the potential outcome:

$$\begin{aligned}
\Big(\pi_{j=0|d=1}^{(0)}, \pi_{j=1|d=1}^{(0)}, \pi_{j=2|d=1}^{(0)}\Big) &= (0.3, 0.5, 0.2) \\
\Big(\pi_{j=0|d=1}^{(1)}, \pi_{j=1|d=1}^{(1)}, \pi_{j=2|d=1}^{(1)}\Big) &= (0.2, 0.5, 0.3) \\
\Big(\pi_{j=0|d=0}^{(0)}, \pi_{j=1|d=0}^{(0)}, \pi_{j=2|d=0}^{(0)}\Big) &= (0.2, 0.5, 0.3) \\
\Big(\pi_{j=0|d=0}^{(1)}, \pi_{j=1|d=0}^{(1)}, \pi_{j=2|d=0}^{(1)}\Big) &= (0.2, 0.4, 0.4)
\end{aligned}$$

Under this DGP, PT1 holds since

$$\begin{aligned}
&\mathbb{E}[\widetilde{Y}_{i1}(0) - \widetilde{Y}_{i0}(0) \mid D_i = 1] - \mathbb{E}[\widetilde{Y}_{i1}(0) - \widetilde{Y}_{i0}(0) \mid D_i = 0] \\
&= [\pi^{(1)}_{2|1} - \pi^{(0)}_{2|1}] - [\pi^{(1)}_{2|0} - \pi^{(0)}_{2|0}] \\
&= 0.1 - 0.1 = 0.
\end{aligned}$$

However, PT2 does not hold because

$$\begin{aligned}
&\mathbb{E}[\check{Y}_{i1}(0) - \check{Y}_{i0}(0) \mid D_i = 1] - \mathbb{E}[\check{Y}_{i1}(0) - \check{Y}_{i0}(0) \mid D_i = 0] \\
&= \Big\{[\pi^{(1)}_{2|1} + \pi^{(1)}_{1|1}] - [\pi^{(0)}_{2|1} + \pi^{(0)}_{1|1}]\Big\} - \Big\{[\pi^{(1)}_{2|0} + \pi^{(1)}_{1|0}] - [\pi^{(0)}_{2|0} + \pi^{(0)}_{1|0}]\Big\} \\
&= \{(0.3 + 0.5) - (0.2 + 0.5)\} - \{(0.4 + 0.4) - (0.3 + 0.5)\} = 0.1.
\end{aligned}$$

Thus, this example demonstrate that with the same data, Δ_2 can be consistently estimated with $\widetilde{Y}_{it}$ but Δ_1 cannot be estimated without bias, even though we have the same data generating process behind the two transformations. Obviously, it is also trivial to construct an example where PT2 holds but PT1 does not.

D Additional Empirical Results

D.1 Different coding of partisanship

Given the partisan nature of the gun control policies, it is important to understand if the effect of mass shootings could differ by respondents' party identification. The coding of partisanship, however, slightly different across studies. In the main text, I relied on the 3-point scale party identification question (`pid3`), which asks respondents the following question,

```
Generally speaking, do you think of yourself as a ...?
  (1) Democrat; (2) Republican; (3) Independent;
  (4) Other; (5) Not sure; (8) Skipped.
```

In the analysis, I exclude respondents who do not select option `(1)`, `(2)` or `(3)`.

In the survey, respondents are also asked to place themselves on a granular scale of partisanship (`pid7`):

```
 Would you call yourself a strong Democrat or a not very strong Democrat?  Would
you call yourself a strong Republican or a not very strong Republican?  Do
you think of yourself as closer to the Democratic or the Republican Party?


  (1) Strong Democrat; (2) Not very strong Democrat; (3) Lean Democrat;
  (4) Independent; (5) Lean Republican; (6) Not very strong Republican;
```

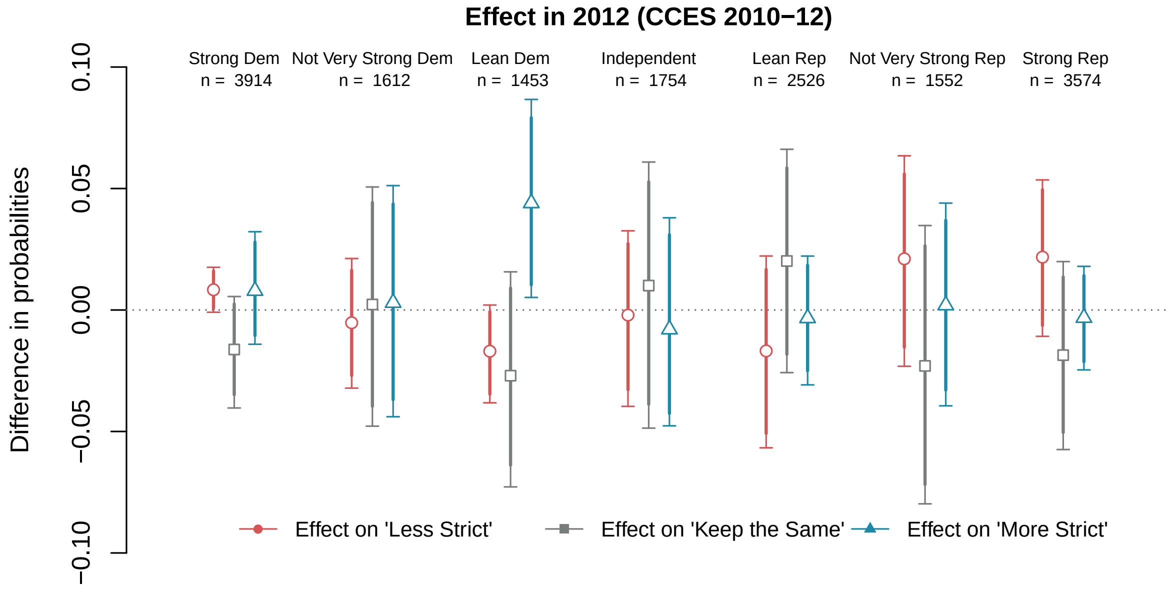

Figure D.1: Estimated effect based on a 7-point scale partisanship (`pid7`).

```
    (7) Strong Republican; (8) Not sure; (98) Skipped.
```

Figure D.1 shows the result of the analyses that use `pid7` to construct partisan subgroups. We can see that the effect is concentrated among "Lean Democrats."

On the other hand, Barney and Schaffner (2019) constructs the partisanship variable based on `pid7` but collapses it into a 3-point scale, "Democrat", "Independent" and "Republican". The major difference from the self-reported `pid3` is that "leaners" are classified as partisans (i.e., not independents). Figure D.2 reports the estimate based on the partisan coding of Barney and Schaffner (2019).

D.2 Different estimands

Figure D.3 shows estimated bounds on $\tau = \Pr(Y_{i1}(1) \geq Y_{i1}(0) \mid D_i = 1)$ (gray lines) and $\eta = \Pr(Y_{i1}(1) > Y_{i1}(0) \mid D_i = 1)$ (red lines). The explicit formula of the bound for each estimand is given in Section B.1. I find that bounds for the two estimands diverge suggesting there are many observations that have $Y_{i1}(1) = Y_{i1}(0)$ in the population. For example, among Democrats, the bound for τ is between around 0.8 and 1.0 which might suggest that proportion of Democrats who supports gun control when treated is extremely high. However, the bound for η is between around 0.2 and 0 which might suggest that proportion of Democrats who have *strictly* prefer a more strict gun control is very small. This is a problem discussed in Lu et al. (2018) where τ and η cannot be informative when there are many units who does not change attitudes by the treatment. Therefore, it appears that we need to turn to other estimators that avoid this problem (e.g., Chiba, 2017; Lu, 2018).

D.3 Different treatment cutoff

Although Newman and Hartman (2019) define the exposure by the 100-mile cutoff, Barney and Schaffner (2019) consider different threshold to assess the robustness of the results. Following their analysis, I consider an alternative threshold of 25 miles. Figure D.4 shows the result. Note that the "prior exposure" is also defined by the 25-mile cutoff.

D.4 Two year subset of three-wave panel

In this section, I present two additional results based on the two-wave panel from CCES. Figure D.5 shows estimated effects based on the sub-group analysis taking interactions between the prior exposure variable and partisan identification. In the figure, circles (triangles) show estimates of Δ_2 (Δ_1) and thin (thick) line indicate 90% (95%) confidence intervals. Confidence intervals are computed based on block bootstraps (blocked at the zip code level). I find that among no-prior-exposure group, the effect is concentrated among Democrats (Δ_2 for Democrats is estimated positive and statistically different from zero at the 10% level, while Δ_1 is not statistically significant). On the other hand, effects for Independents and Republicans are both indistinguishable from zero at the 10% level (neither Δ_2 nor Δ_1). I also find that effects are almost zero in the prior-exposure groups regardless of partisanship.

E Details of the Application

A list of method used in the original studies Table 1 summarizes the methods used in the original papers.

Table 1: Methodologies used in the original studies. Abbreviation: Newman and Hartman (2019) as NH19, Barney and Schaffner (2019) as BS19 and Hartman and Newman (2019) as HN19.

	NH19	BS19	HN19
`ordered logit (RE)`	✓(with Lag DV)	✓	✓
`ordered logit (FE)`			✓
`linear two-way FE`		✓	

Coding of mass shootings Newman and Hartman (2019) uses the following criteria to determine if an incident constitutes a mass public shooting: "(1) firearms as the primary weapon used, (2) attacks on non-family members of the general public and (3) attacks in which at least three or more individuals were injured or killed." (Newman and Hartman, 2019, p.8). See the original studies for the detail of why these criteria are selected. Note that the definition of the "treatment" is slightly different between Newman and Hartman (2019) and Barney and Schaffner (2019). I follow the definition used by Barney and Schaffner (2019); Please see Barney and Schaffner (2019) for the discussion on this point.

Survey outcome The ordering of the response categories is not exactly the same as the original question in CCES 2010–2012 panel. Originally in the survey, the choices are given as (1) More Strict; (2) Less Strict; (3) Kept As They Are (please see `CC10_320` and `CC12_320` in "Guide to the 2010-12 CCES Panel Study" available at `https://doi.org/10.7910/DVN/24416/79YKV2`). In the main text, I follow the coding of Newman and Hartman (2019) and Barney and Schaffner (2019) and treat "Kept As They Are" as the middle category.

Figure E.1 shows the distribution of the outcome in 2010 (top) and 2012 (bottom) where the blue bars correspond to the treatment group and the gray bars correspond to the control group.

F Simulation Studies

In this section, I present two Monte Carlo studies to investigate finite sample performances of the proposed method. The first simulation assesses performance of the proposed estimator for the causal effect where I compare the proposed estimator against the standard difference-in-differences with dichotomized outcome and the ordered probit regression. The result shows that the proposed estimator is unbiased to the causal effects and the confidence interval has nominal coverage, while the other two method are biased and thus the confidence intervals fail to maintain the coverage. The second simulation studies the finite sample performance of the proposed procedure for the diagnostic in Section 3.4. I demonstrate that the type I error is controlled under the range of equivalence threshold that is compatible with the null and that the power converges to one when the equivalence threshold is chosen reasonably.

F.1 Estimating causal effects

In this first simulation study, I investigate the finite sample performance of the proposed estimator under the correct model specification. The potential outcome is generated by first drawing the latent utilities from the normal distribution. For the potential outcome under the control, the following set of parameters are used to generate the data: $\theta_{00} = (-0.5, 1.5)^\top$, $\theta_{01} = (1, 1)^\top$ and $\theta_{10} = (-1.5, 2)^\top$. The parameters for the counterfactual outcome θ_{11} is set according to the identification formula in Proposition 1. The parameters for generating the potential outcome under the treatment, $Y_{i1}(1)$, are set as $\mu = 1.5$ and $\sigma = 1.5$.

After generating the latent utilizes, they are transformed into categorical outcome with J categories based on the set of cutoffs. In this simulation, I consider $J \in \{3, 5, 7\}$ and also I vary the number of units $n \in \{1000, 2500, 5000\}$.

Since there are $J-1$ possible treatment effects to consider, that is, $\{\Delta_j\}_{j=1}^{J-1}$, estimators are evaluated on averaging the loss over $J-1$ treatment effect estimates. Specifically, I consider the following metrics:

$$\begin{aligned}
\overline{\mathsf{Abs.Bias}} &= \frac{1}{(J-1)} \sum_{j=1}^{J-1} \left| \frac{1}{S} \sum_{s=1}^{S} (\widehat{\Delta}_j^{(s)} - \Delta_j) \right| \\
\overline{\mathsf{RMSE}} &= \frac{1}{(J-1)} \sum_{j=1}^{J-1} \left\{ \frac{1}{S} \sum_{s=1}^{S} (\widehat{\Delta}_j^{(s)} - \Delta_j)^2 \right\}^{1/2} \\
\overline{\mathsf{Coverage}} &= \frac{1}{(J-1)S} \sum_{j=1}^{J-1} \sum_{s=1}^{S} \mathbf{1}\Big\{ \Delta_j \in \widehat{C}_{j,1-\alpha/2}^{(s)} \Big\}
\end{aligned}$$

where $\widehat{\Delta}_j^{(s)}$ is the estimate of Δ_j under sth Monte Carlo iteration and $\widehat{C}_{j,1-\alpha/2}^{(s)}$ is the $100 \times (1-\alpha/2)\%$ confidence interval for Δ_j.

Figure F.1 shows the result. Left panel shows the absolute bias of the estimate. We see that the bias is larger when the sample is relatively small for $J=5$ as the number of observations in each category tend to be smaller. However, in general, estimates are unbiased. Middle panel shows the RMSE. It shows that RMSE decreases as the sample size increases and the variance is smaller when the number of categories are smaller. Finally, the right panel shows the coverage of 90% confidence intervals. We can see that for both cases, confidence intervals have nominal coverage regardless of sample size.

F.2 Testing procedure

In this section, I investigate a finite sample performance of the proposed testing procedure. Specifically, I conduct a Monte Carlo simulation with a scenario that Assumption 3 is violated in the pre-treatment periods. Outcomes are generated first by simulating the latent utilities. The latent utilities are simulated according to the normal distribution with mean μ_{dt} and variance σ_{dt}^2,

$$Y_{dt}^* \sim \mathcal{N}(\mu_{dt}, \sigma_{dt}^2) \tag{F.1}$$

I set $\theta_{00} = (-0.5, 1.5)^\top$, $\theta_{01} = (1,1)^\top$, $\theta_{10} = (-1.5, 2)^\top$ and $\theta_{11} = (1.5, 1.5)^\top$. This parameter specification leads to the true maximum deviation $t_{\max} \approx 1.4$. Clearly, this does not satisfy Assumption 3 which requires $t_{\max} = 0$.

After simulating Y_{dt}^*, categorial outcomes are generated based on cutoffs κ. In this simulation, I consider three cases: $J \in \{3, 5, 7\}$. For $J=3$ and $J=5$, the same cutoffs as in the previous simulation are used. For $J=7$, I use $\kappa = (-0.5, -0.2, 0.1, 0.4, 0.7, 1.0)^\top$.

In order to assess how the test performs depending on a choice of equivalence thresholds, I vary δ. The value of δ is chosen such that in some range of δ, the null of $t_{\max} \geq \delta$ is true and in other range of δ the null is false (i.e., $t_{\max} < \delta$). For the range of δ that satisfies $t_{\max} \geq \delta$, I set $\delta \in \{t_{\max} - 0.05, t_{\max} - 0.01, t_{\max}\}$. We would expect that the test is more likely to reject the alternative when $\delta = t_{\max} - 0.05$. For the range of δ that does not satisfy $t_{\max} \geq \delta$, I set $\delta \in \{t_{\max} + 0.05, t_{\max} + 0.01, t_{\max} + 0.10\}$. Among them, we expect that the test can reject the null most likely when $\delta = t_{\max} + 0.10$.

Figure F.2 shows the results for this simulation. The upper panels show type I errors when the choice of δ is consistent with the data (i.e., $t_{\max} \geq \delta$). Recall that the data is simulated such that the equivalence does not hold. Thus, we would expect that the null, $H_0\colon t_{\max} \geq \delta$, is not rejected and the probability of falsely rejecting the null (type I error) should be less than α. In fact, we can see that the proposed testing procedure controls the type I error. In addition, the smaller value of δ (i.e., a smaller rejection region for H_0) leads to lower type I error. The lower panels show results for type II errors when the choice of δ is not consistent with the data (i.e., H_0 is false). We can see that the test struggles to reject the null when δ is set to close to $t_{\max}$. When δ is set to a value far away from $t_{\max}$, type II error converges to zero as sample size increases.

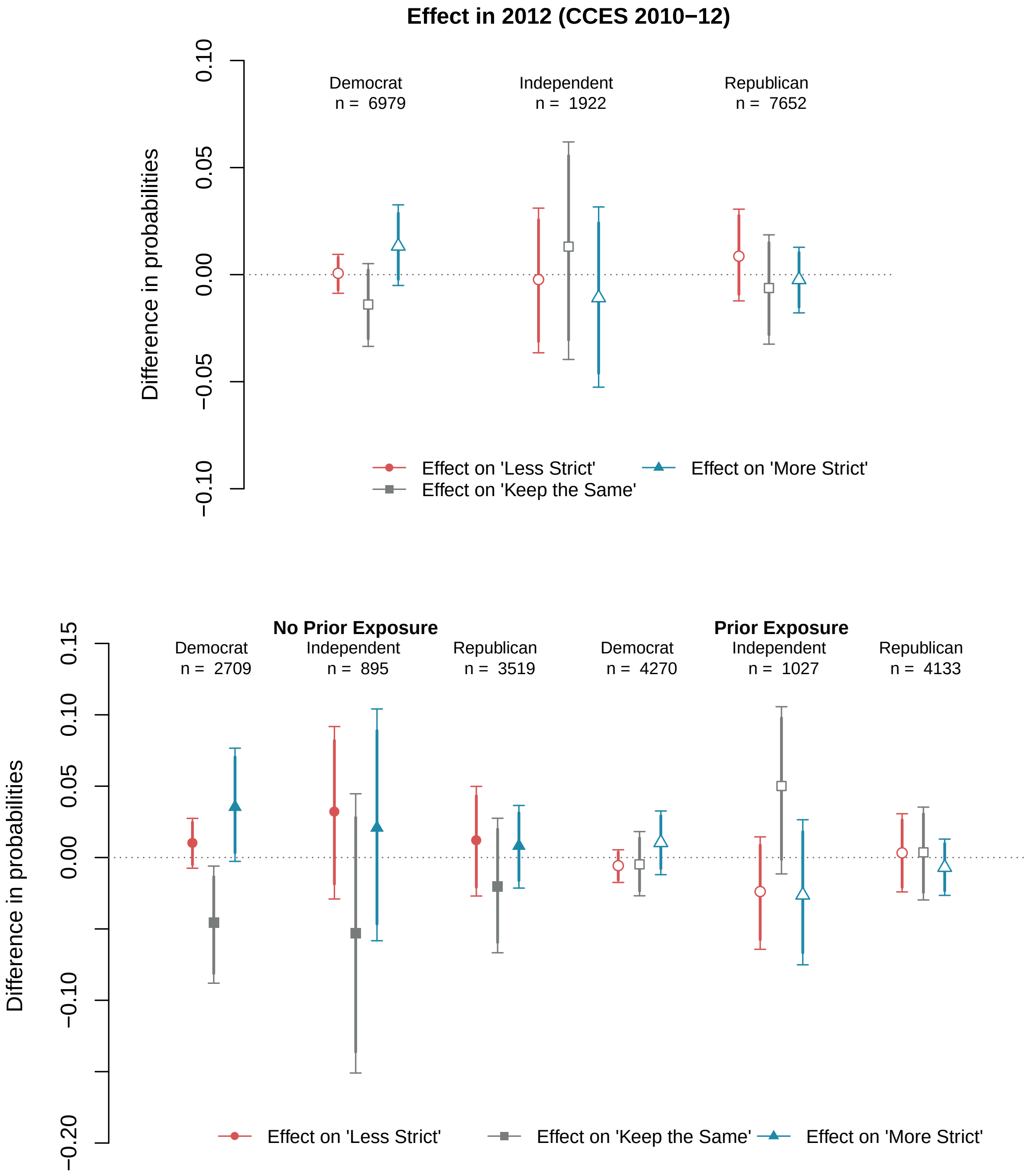

Figure D.2: Estimated effects based on partisanship based on a coding used in Barney and Schaffner (2019).

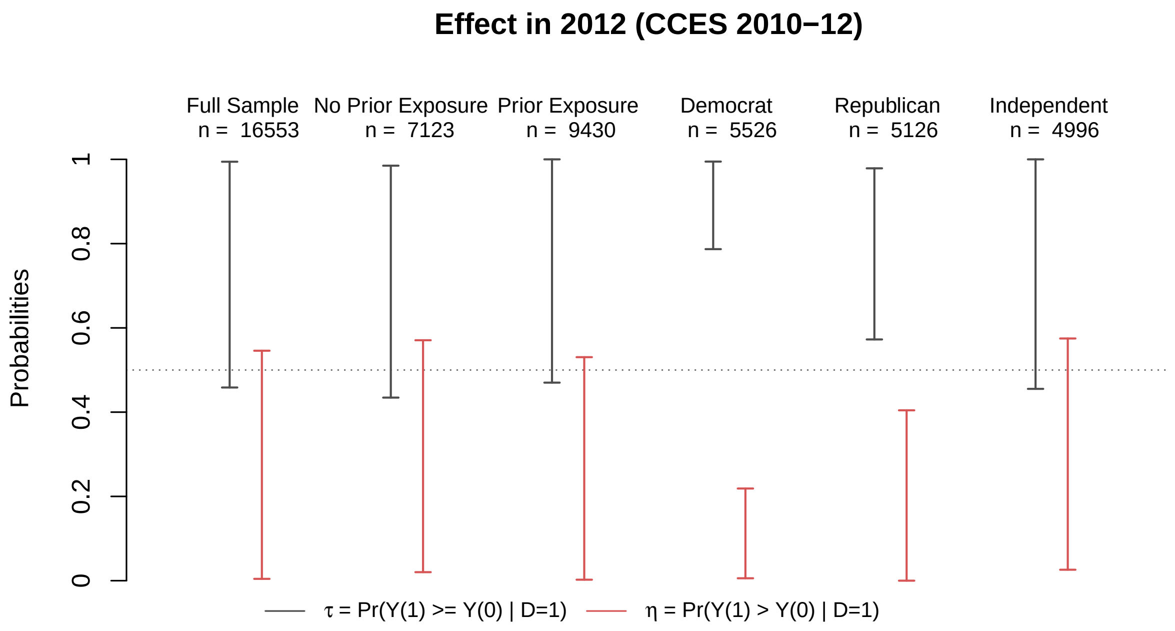

Figure D.3: Estimated bound for τ (gray lines) and η (red lines).

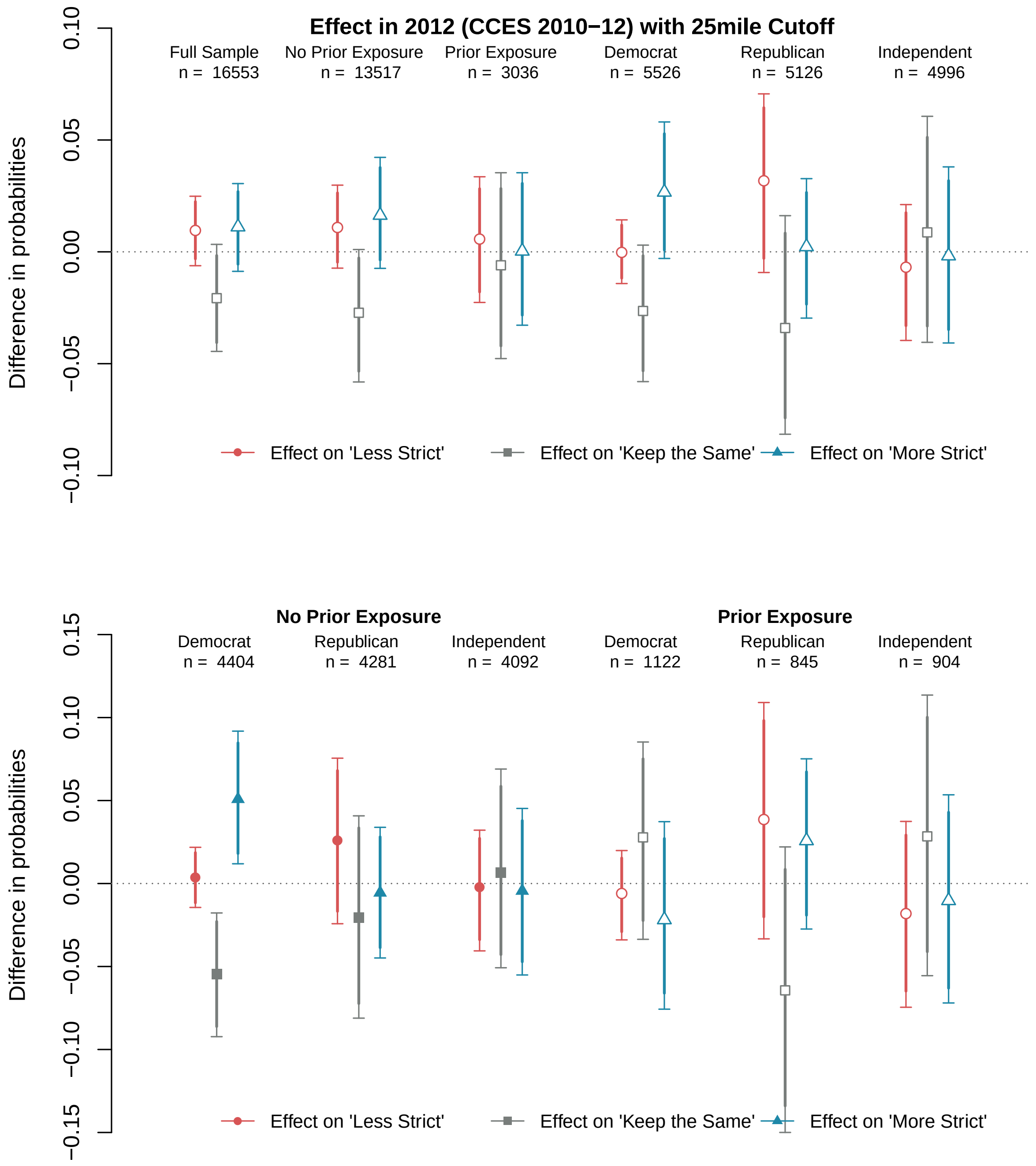

Figure D.4: Estimated causal effect with **25-mile** cutoff as the threshold for the exposure. Circles are the estimate of ζ_0, square are the estimate of ζ_1 and triangles are the estimate of ζ_2. Thin (thick) lines indicate 90% (95%) confidence intervals.

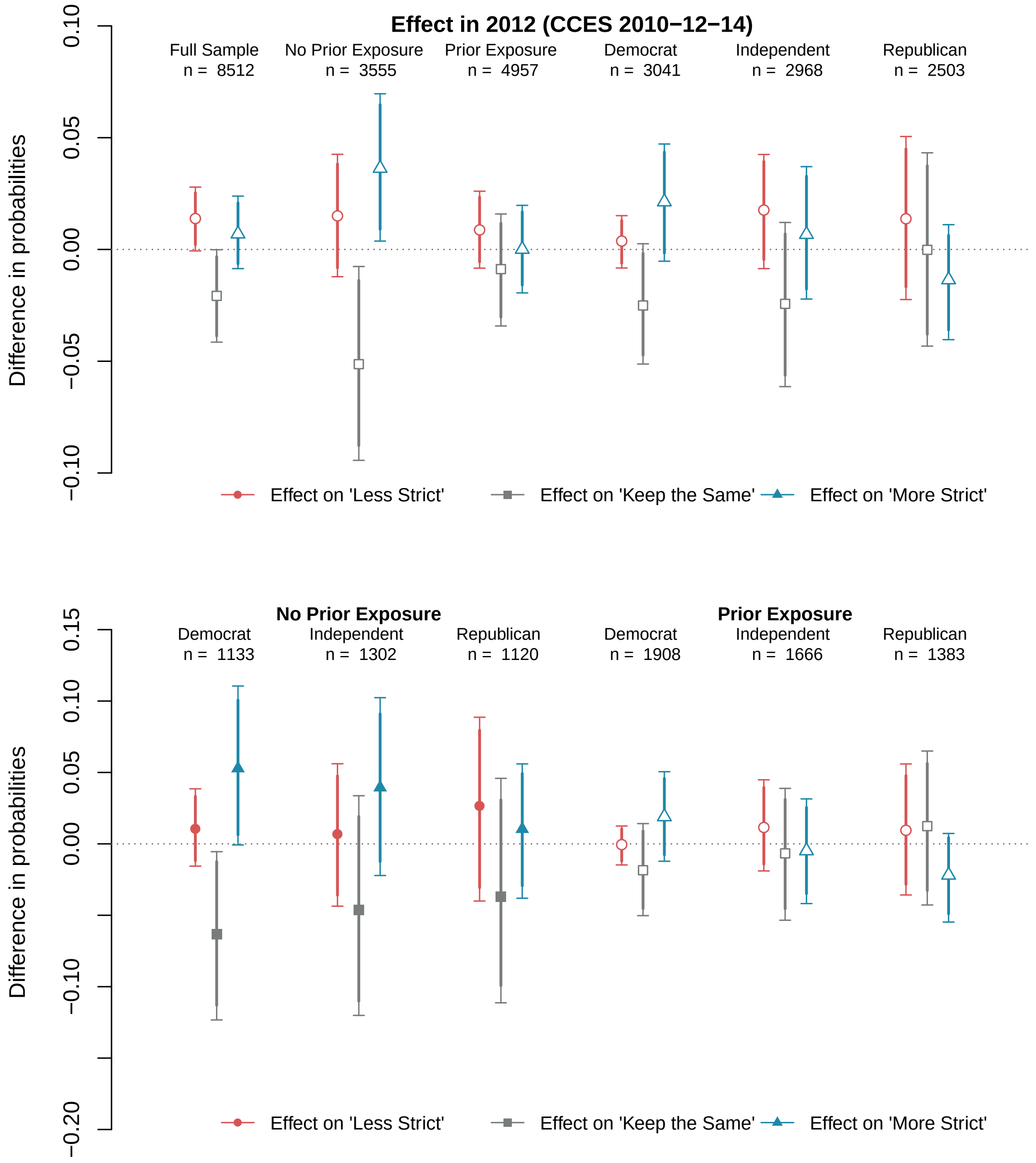

Figure D.5: Estimated effects using the two-year subset of the three-year panel. Circles are the estimate of ζ_0, square are the estimate of ζ_1 and triangles are the estimate of ζ_2. Thin (thick) lines indicate 90% (95%) confidence intervals.

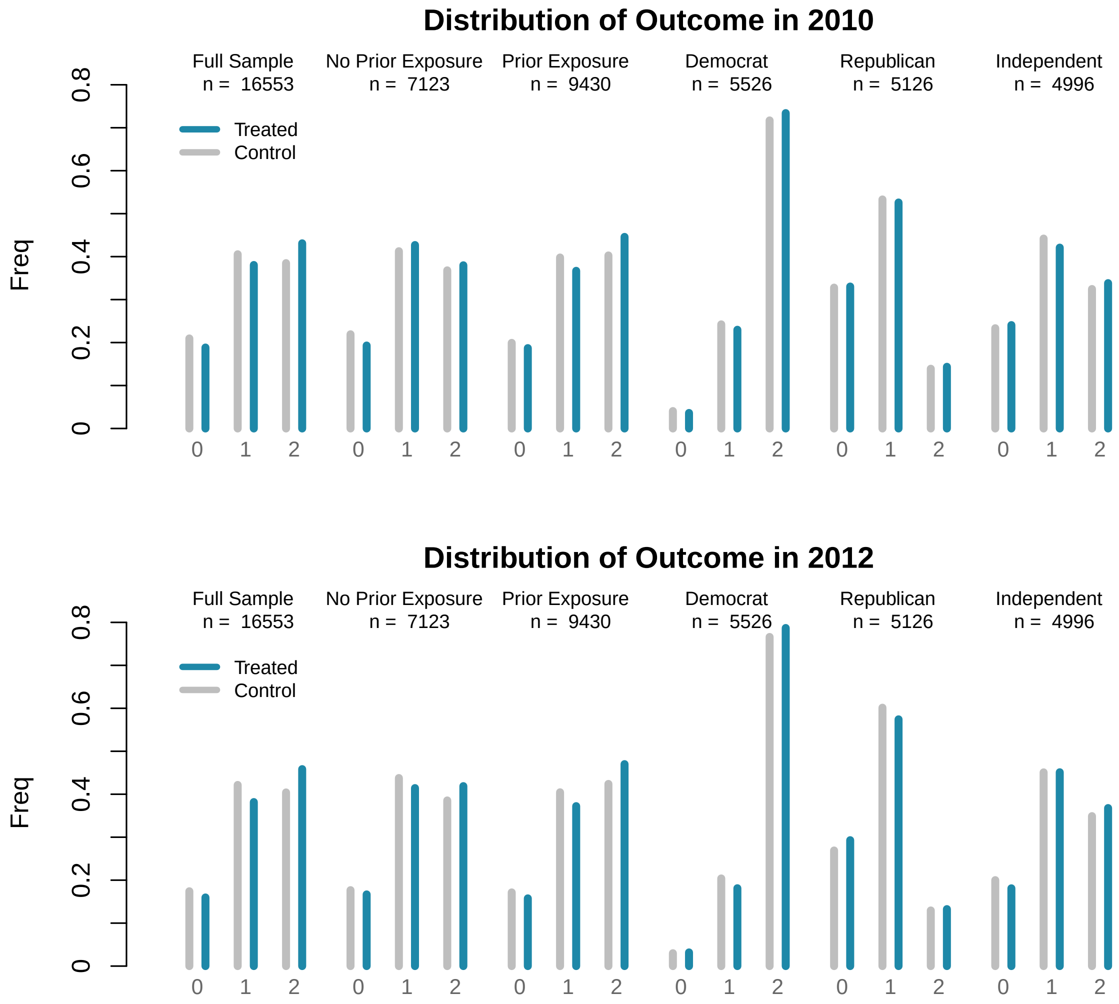

Figure E.1: Distribution of outcomes: `(0): less-strict`, `(1): kept-as-they-are` and `(2): more-strict`. The top panel shows the distribution of 2010 and the bottom panel is for 2012. Bars in blue (gray) shows distributions for the treated (control) group.

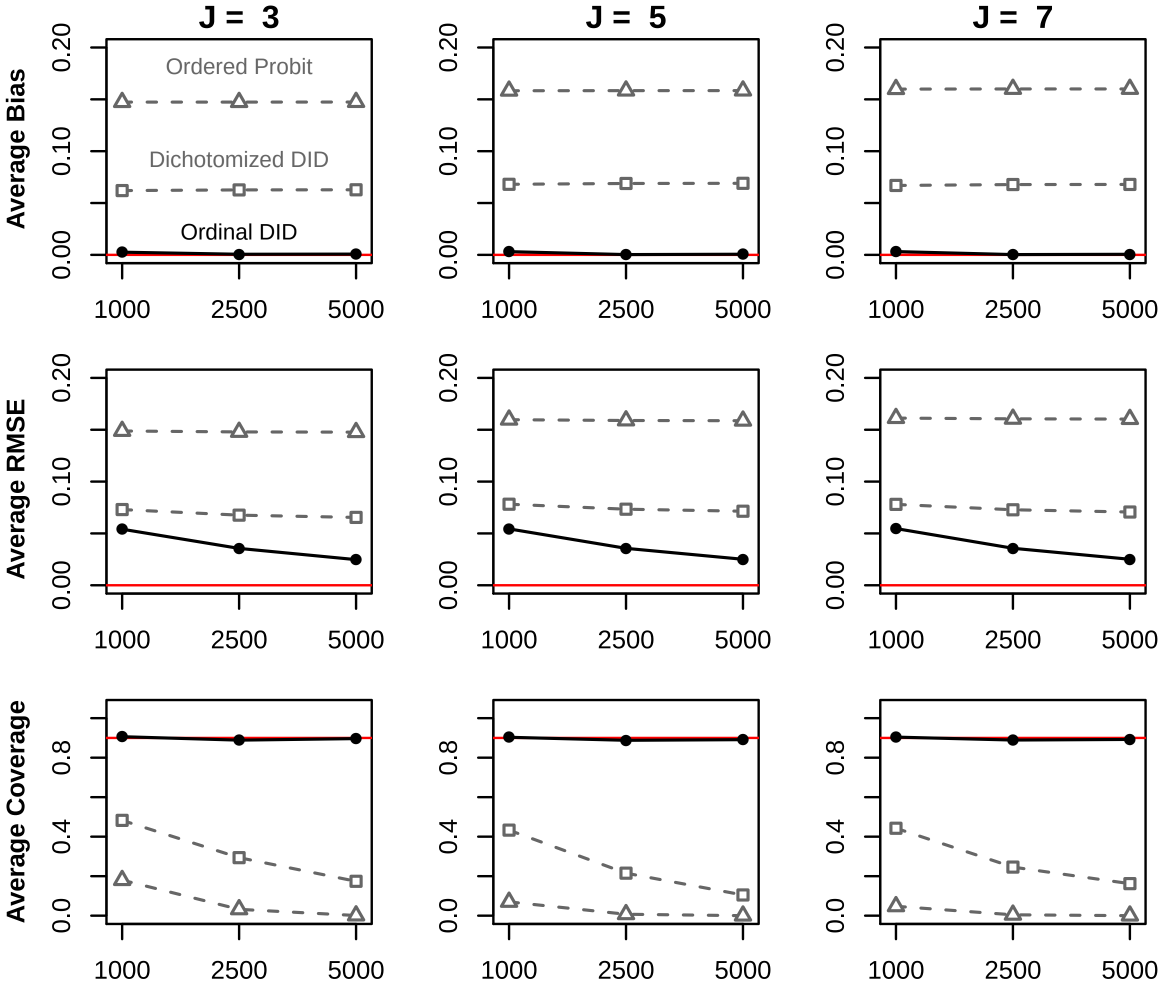

Figure F.1: Simulation Results. Top row: Absolute bias ($\overline{\mathsf{Abs.\ Bias}}$). Middle: RMSE ($\overline{\mathsf{RMSE}}$). Bottom: Coverage based on the 90% confidence interval ($\overline{\mathsf{Coverage}}$). As expected from the general result of Maximum Likelihood, the estimate is unbiased and the confidence interval maintains nominal coverage under the correct specification.

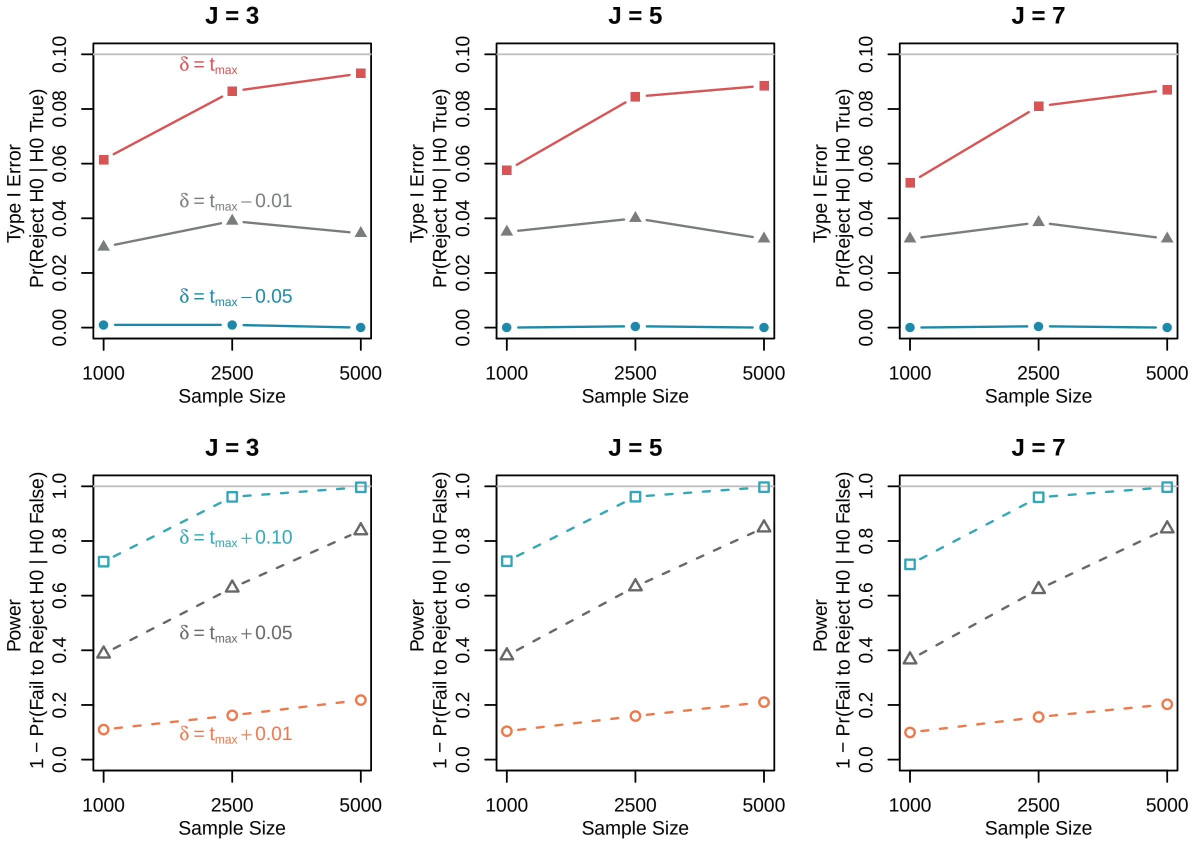

Figure F.2: Finite sample performance of the proposed testing procedure: Type I error (upper panel) and power (lower panel). The upper panel shows rejection probabilities of H_0 under thresholds that are compatible with H_0 (i.e., H_0 is true: $t_{\max} \geq \delta$ holds). The lower panel shows the power curve in a range of δ that is not compatible with H_0 (i.e., H_1 is true: $t_{\max} < \delta$).