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

Soichiro Yamauchi†

This version: February 27, 2026
First draft: October 29, 2019

Abstract

Difference-in-differences (DID) is widely used to estimate causal effects with repeated observations. The standard DID relies on the parallel trends assumption that requires the linearity in the potential outcome. However, social science research heavily relies on ordinal outcomes, such as Likert-type survey responses, where the linearity assumption is not appropriate. This paper develops a methodology for conducting DID analysis for ordinal outcomes. The proposed method allows scholars to estimate causal estimands that are suitable for ordinal outcomes without imposing the linearity or outcome transformation assumption. Building on the discrete choice literature that models ordinal outcomes using a latent continuous variable framework, I develop a set of identification assumptions. The paper also develops an equivalence testing procedure to assess the identification assumption with pre-treatment data. I extend the method to a semiparametric setting where a rich set of covariates is available, and to a staggered adoption design when treatment timing varies across units. The proposed method is applied to revisit studies on estimating the causal effect of mass shooting incidents on the public's support for gun control.

Keywords: Difference-in-differences, equivalence testing, ordinal outcome, panel data

*I am grateful to Matt Blackwell, Gary King, Kosuke Imai, Molly Offer-Westort, Ikuma Ogura, Shun Yamaya, 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. The R package `orddid` is available for implementing the proposed methodology at https://github.com/soichiroy/orddid.

†Assistant Professor, Department of Political Science, University of California, San Diego. Email: soichiro@ucsd.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 causal effects 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; Callaway and Sant'Anna, 2021; Chang, 2020; Park and Tchetgen Tchetgen, 2022; Liu, Wang and Xu, 2024), most of the existing methods identify and estimate the treatment effect under the linearity assumption. The parallel-trends assumption imposes a restriction on the potential outcomes such that the mean of the treatment and the control group has identical trends in the absence of the treatment. Therefore, the assumption is meaningful only when the difference between two potential outcomes is well defined, as for continuous outcomes.

In social science research, however, many outcomes of interest are measured on an ordinal scale. For example, in political science, scholars measure voters' ideology on a scale from "very liberal" to "very conservative" or measure attitudes toward a policy item from "strongly disagree" to "strongly agree." In fact, due to space limitations and other administrative reasons, most of the questions asked in major social science surveys are ordinal. When the outcome is measured on such a scale, it is difficult to define the "mean" of non-numeric variables, and thus the usual definition of the treatment effect as the difference between two potential outcomes is not well defined (e.g., Volfovsky, Airoldi and Rubin, 2015; Lu, Ding and Dasgupta, 2018), unless strong assumptions about the scale are imposed. This implies that the application of the standard DID requires additional restrictions that are not necessarily appropriate. Dichotomizing the outcome is another common practice to apply the standard DID method. Recent results show that, only under very restrictive conditions, parallel trends are invariant to transformations (Roth and Sant'Anna, 2023). This implies that different dichotomizations may lead to different conclusions, not to mention that it requires separate parallel trends assumptions that are not necessarily compatible with each other. Existing DID methods thus require the researcher either to impose linearity on an inherently non-linear scale or to discard ordinal information through dichotomization.

To address these limitations, this paper develops a methodology for estimating causal effects for ordinal outcomes with repeated observations. In contrast to the existing approaches, the proposed method utilizes the ordering of categories through a latent variable framework, building on the formulation for ordinal outcomes in discrete choice models (e.g., Maddala, 1983; Train, 2009). I propose a set of identification assumptions that extend the distributional DID framework of Athey and Imbens (2006) to ordinal outcomes (Section 2). Instead of assuming linearity on the actual outcome, I impose identification assumptions on the latent continuous variable underlying

the ordinal outcome. This allows for the identification of the entire distribution of the latent variable for the counterfactual outcome and further allows researchers to estimate various causal quantities for ordinal outcomes, such as the distributional effect and the relative effect (Lu, Ding and Dasgupta, 2018; Lu, Zhang and Ding, 2020).

I also develop an equivalence-based testing procedure that formally assesses the plausibility of the key identification assumption using pre-treatment periods (Section 3). As in the standard DID, the proposed procedure tests whether the assumption holds at least during the pre-treatment periods. I show that the test statistic in the proposed equivalence test is related to the worst-case bias of the treatment effects. This enables scholars to calibrate the equivalence threshold based on the maximum tolerable bias in their applications, and thus overcome the common challenge of equivalence tests regarding the selection of the equivalence threshold. I extend the proposed framework to a semiparametric setting that relaxes the parametric assumption on the latent variable distribution, and to the staggered adoption design where treatment timing varies across units (Section 4).

The proposed methodology is closely related to the broader literature on non-linear DID (e.g., Athey and Imbens, 2006; Sofer et al., 2016; Callaway, Li and Oka, 2018; Glynn and Ichino, 2019; Park and Tchetgen Tchetgen, 2022). Athey and Imbens (2006) extend their method to binary and count outcomes, but provide only partial identification for non-continuous outcomes. Park and Tchetgen Tchetgen (2022) propose an odds-ratio-based assumption applicable to any discrete outcome (also see Richardson, Ye and Tchetgen Tchetgen, 2023; Tchetgen Tchetgen, Park and Richardson, 2024), but their approach treats categories as unordered. Neither considers equivalence-based testing procedures to assess identification assumptions.

Section 5 revisits a recent debate on the effect of mass shootings on attitudes towards gun control regulations. Hartman and Newman (2019) and Newman and Hartman (2019) argue that mass shootings increase support for stricter gun control, while Barney and Schaffner (2019) find no strong evidence for this claim. Using the proposed methodology, I re-analyze a panel of over 16,000 respondents and find little evidence that mass shootings affect the support for gun control. An open-source `R` package, `orddid`, implements the proposed methodology and is available at `https://github.com/soichiroy/orddid`.

2 The Proposed Methodology

2.1 The setup and estimands

Let $Y_{it} \in \{0, \ldots, J-1\} \equiv \mathcal{J}$ denote the observed outcome measured on an ordinal scale with J categories ($J \geq 3$) for unit $i \in \{1, \ldots, n\}$ and time $t \in \{0, 1\}$. As a running example, I consider the

effect of mass shootings on attitudes toward gun control regulations (Section 5), where $J = 3$ and the outcome is coded as $Y = 0$ (less strict), $Y = 1$ (kept as they are), and $Y = 2$ (more strict). The binary treatment, denoted by $D_i \in \{0, 1\}$, is assigned after Y_{i0} is observed but before time $t = 1$. I 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).

This paper considers two types of causal estimands for the treated units. First, following the conventional practice in the literature, I consider the treatment effect that is defined as the contrast of the two marginal probabilities of the potential outcomes. However, following the growing literature on causal inference with ordinal outcomes (e.g., Volfovsky, Airoldi and Rubin, 2015; Chiba, 2017; Lu, 2018; Lu, Ding and Dasgupta, 2018; Lu, Zhang and Ding, 2020), I consider an estimand that is more suitable for ordinal outcomes.

Conventional estimands As the conventional estimands, I focus on the difference in probabilities per category. Specifically, the effect ζ_j is defined as the difference in probabilities of category j under two conditions,

$$\zeta_j = \mathbb{P}(Y_{i1}(1) = j \mid D_i = 1) - \mathbb{P}(Y_{i1}(0) = j \mid D_i = 1). \tag{2.1}$$

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 the "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 useful to summarize the effects using the cumulative effect Δ_j,

$$\Delta_j = \mathbb{P}(Y_{i1}(1) \geq j \mid D_i = 1) - \mathbb{P}(Y_{i1}(0) \geq j \mid D_i = 1) = \sum_{\ell=j}^{J-1} \zeta_\ell$$

for $j \in \mathcal{J} \backslash \{0\}$. Since Δ_j is a partial sum of ζ_ℓ, identifying ζ_j suffices for all cumulative effects.

Estimand for ordinal outcomes The conventional estimands ζ_j and Δ_j do not fully capture the nature of ordinal outcomes. For example, ζ_j has a natural constraint $\sum_{j=0}^{J-1} \zeta_j = 0$, which means that a positive effect on one category must be compensated by negative effects on other categories. In addition, there would be $J - 1$ number of effects to report for the case of ζ, which makes it difficult to summarize the overall impact of the treatment on the ordinal outcome, especially when

the number of categories increases.

Instead, I consider the relative effect for the treated that is useful to summarize the impact on the ordinal outcome:

$$\tau = \mathbb{P}(Y_{i1}(1) > Y_{i1}(0) \mid D_i = 1) - \mathbb{P}(Y_{i1}(1) < Y_{i1}(0) \mid D_i = 1). \tag{2.2}$$

The relative effect τ captures the overall tendency of the treatment effect on the ordinal outcome. The positive impact on τ indicates that the treatment tends to increase the outcome for the treated units. In our application, $\tau > 0$ implies that the mass shootings generally make people prefer stricter gun control policies.

The benefit of the relative effect τ is that it summarizes the impact on the entire distribution on the outcome, and therefore does not depend on the choice of reference category. The downside of the effect is that it depends on the joint distribution of the potential outcomes, which is not point identified in general unless strong assumptions are imposed on the joint distribution on the potential outcomes. I leverage the sharp bound derived in Lu, Zhang and Ding (2020) to provide the identification bound on τ in the current setting. The upper bound is given by

$$\tau \leq \tau_U = 1 + \min_{1 \leq j \leq J-1} \min_{1 \leq j' \leq J-j} \{\Delta_j + \Delta_{j+j'}\}$$

where τ_U is a function of marginal distributions of the potential outcomes only. The expression of the lower bound and the derivation is presented in Appendix A.1,

2.2 Identification

Since $Y_{i1}(0)$ is not observed for the treated units, the goal is to identify $\mathbb{P}(Y_{i1}(0) = j \mid D_i = 1)$ for all $j \in \mathcal{J}$. 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. While we observe Y_{00}, Y_{01} and Y_{10}, the counterfactual outcome $Y_{11} \sim Y_{i1}(0) \mid D_i = 1$ is not observed.

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 $Y^*_{dt} \in \mathbb{R}$ for all (d,t), and

$$Y_{dt} = \begin{cases} 0 & \text{if} \quad \kappa_0 \leq Y^*_{dt} < \kappa_1 \\ j & \text{if} \quad \kappa_j \leq Y^*_{dt} < \kappa_{j+1} \\ J-1 & \text{if} \quad \kappa_{J-1} \leq Y^*_{dt} \leq \kappa_J \end{cases} \tag{2.3}$$

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 a latent potential outcome Y^*_{dt} defined on a continuous space. The assumption allows us to work on a continuous space through Y^*_{dt}. In the gun control application, Y^*_{dt} represents the continuous latent attitude toward firearm regulation, and respondents report one of the three categories depending on where this latent attitude falls relative to the cutoffs. The cutoff values are constant across groups and time, but this is not restrictive because any group and time specific cutoffs can be absorbed into the distribution of Y^*_{dt} (e.g., by shifting mean and variance).

Since $\{\zeta_j\}_{j \in \mathcal{J}}$ depends on the entire marginal distribution of the potential outcome, I further impose a distributional assumption on Y^*_{dt}.

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

$$Y^*_{dt} \sim \mu_{dt} + \sigma_{dt} U, \quad U \sim F \tag{2.4}$$

where μ_{dt} is the location, σ_{dt} is the scale parameter, and F is the base distribution.

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 latent outcomes between time and groups are different up to mean and the scale.

The assumption is general in that it can accommodate a wide class of distributions as the base distribution F. For example, in the standard ordinal regression analysis, the normal distribution or the logistic distribution is often assumed as the base distribution. However, other parametric distributions such as the t-distribution or the extreme value distribution, or some mixture distributions can also be used as the base distribution. Note that the joint distribution of the latent utilities is left unspecified, so units can have correlated latent utilities over time.

One limitation of Assumption 2 is that a researcher must specify the base distribution F in advance. A misspecification of the base distribution may lead to biased estimates of the treatment

effects. To overcome the limitation, in Section 4.1, I develop a semiparametric approach that allows for F to be unknown and estimated nonparametrically from the data, when a rich set of pre-treatment covariates is available. The semiparametric approach allows scholars to avoid imposing a specific distribution on F for the latent utilities. This section maintains the assumption that the distribution F is appropriately specified.

Finally, I assume that the *shift in the distribution* of latent outcomes across time is constant between the treatment and the control groups. This allows us to use the distributional change in the control group over time to infer the counterfactual distribution for the treated group.

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

$$q_1(v) = q_0(v) \tag{2.5}$$

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. The assumption is originally introduced by Athey and Imbens (2006), and Sofer et al. (2016) calls it the "equi-confounding" assumption. The assumption requires that the unobserved confounders affect the distribution of the latent outcome in the same way across groups. In the gun control application, this means that the distributional shift in latent attitudes between the pre- and post-periods would have been the same for respondents near and far from mass shooting sites, absent the shootings.

Figure 1 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 $d = 0, 1$.

Under Assumption 2, $q_d(v)$ admits a simple form

$$q_d(v) = F_U\left(\frac{\mu_{d1} - \mu_{d0}}{\sigma_{d0}} + \frac{\sigma_{d1}}{\sigma_{d0}} F_U^{-1}(v)\right)$$

where F_U is the CDF of U. This implies that Assumption 3 reduces to the non-linear constraints on the parameters:

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

However, I maintain the original form of the assumption as it is simpler to interpret and visualize. In Section 3, I develop a diagnostic tool that leverages the maximum deviation between $q_1(v)$ and $q_0(v)$ to assess the plausibility of the assumption.

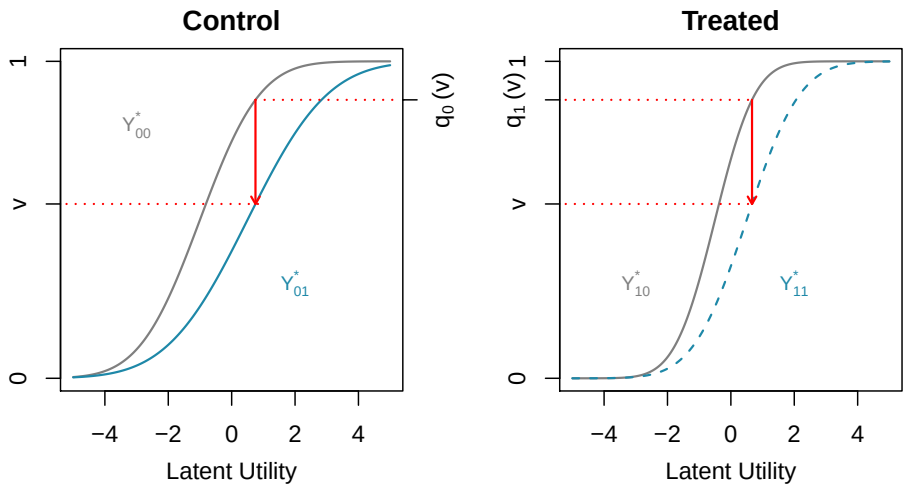

Figure 1: Graphical illustration of Assumption 3. Left (right): cumulative distribution functions of the latent utilities Y^*_{dt} under the control (treatment) condition. Blue (gray) lines indicate the distribution for time $t = 1$ ($t = 0$). Dashed line on the right panel is the distribution of counterfactual 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).

Assumption 3 incorporates the standard parallel trends assumption on the latent outcome as a special case:

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

which is equivalent to $\mu_{11} - \mu_{10} = \mu_{01} - \mu_{00}$. However, this requires that the variance of the latent outcome is constant across time and groups, $\sigma_{dt} = \sigma$ for all d and t. Assumption 3 allows for heterogeneity in the variance, which is more flexible and suitable for the distributional treatment effect.

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, and 3, together with the normalization that $\sigma_{00} = 1$ and $\kappa_1 = 0$, the distribution of the counterfactual latent utility Y^*_{11} is identified as*

$$Y^*_{11} \sim \mu_{11} + \sigma_{11} U \tag{2.6}$$

where

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

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 the base distribution* U*.*

Proof is in Appendix E. 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.

2.3 Estimation and Inference

I propose a three-step estimator to estimate the marginal distribution of the potential outcomes. Once the estimates of the marginal distribution are obtained, the causal effects are estimated by the plug-in method. In Proposition 2, I show that the proposed estimator is consistent and asymptotically normal under the standard regularity conditions for M-estimator.

Estimating the marginal distribution

Step 1: Estimate parameters and cutoffs for the Y_{00}. We estimate the cutoffs and mean parameter, while normalizing the scale parameter to $\sigma_{00} = 1$ and $\kappa_1 = 0$ for identification.

$$(\widehat{\mu}_{00}, \widehat{\boldsymbol{\kappa}}) = \arg\max \sum_{i:D_i=0} \sum_{j \in \mathcal{J}} \mathbf{1}\{Y_{i0} = j\} \log\left\{F(\kappa_{j+1} - \mu_{00}) - F(\kappa_j - \mu_{00})\right\}.$$

Step 2: Estimate parameters for Y_{01} and Y_{10}. We estimate the location and scale parameters for each outcome, while fixing the cutoffs to the estimates obtained in Step 1.

$$(\widehat{\mu}_{01}, \widehat{\sigma}_{01}) = \arg\max \sum_{i:D_i=0} \sum_{j \in \mathcal{J}} \mathbf{1}\{Y_{i1} = j\} \log\left\{F((\widehat{\kappa}_{j+1} - \mu_{01})/\sigma_{01}) - F((\widehat{\kappa}_j - \mu_{01})/\sigma_{01})\right\}$$

Parameters for Y_{10} are estimated similarly.

Step 3: Finally, we recover the distribution of the counterfactual outcome Y_{11} as

$$\widehat{\mathbb{P}}(Y_{11} = j) = F\left(\frac{\widehat{\kappa}_{j+1} - \widehat{\mu}_{11}}{\widehat{\sigma}_{11}}\right) - F\left(\frac{\widehat{\kappa}_j - \widehat{\mu}_{11}}{\widehat{\sigma}_{11}}\right)$$

where $\widehat{\mu}_{11} = \widehat{\mu}_{10} + \widehat{\sigma}_{10}(\widehat{\mu}_{01} - \widehat{\mu}_{00})$ and $\widehat{\sigma}_{11} = \widehat{\sigma}_{10}\widehat{\sigma}_{01}$.

Estimating the causal effects The first estimand $\widehat{\zeta}_j$ can be immediately estimated by plugging in the estimate of the marginal distribution of Y_{11} obtained in the previous step:

$$\widehat{\zeta}_j = \widehat{\mathbb{P}}(Y_{i1} = j \mid D_i = 1) - \widehat{\mathbb{P}}(Y_{11} = j)$$

for all $j \in \mathcal{J}$, where $\widehat{\mathbb{P}}(Y_{i1} = j \mid D_i = 1) = \sum_{i=1}^{n} D_i \mathbf{1}\{Y_{i1} = j\}/n_1$.

The bound on the relative effect τ can be estimated by plugging in the estimates of the marginal distributions of Y_{11} and Y_{01}. Specifically, the upper bound is estimated as

$$\widehat{\tau}_U = 1 + \min_{1 \le j \le J-1} \min_{1 \le j' \le J-j} \{\widehat{\Delta}_j + \widehat{\Delta}_{j+j'}\}$$

where $\Delta_j = \widehat{\mathbb{P}}(Y_{i1} \ge j \mid D_i = 1) - \widehat{\mathbb{P}}(Y_{11} \ge j)$.

The following proposition establishes the consistency and asymptotic normality of the proposed estimator for treatment effects.

Proposition 2 (Asymptotic properties)**.** *Assumption 1, 2, and 3 hold. In addition, the regularity conditions in Assumption A.12 hold. Then, the estimator* $\{\widehat{\zeta}_j\}_{j=0}^{J-1}$ *is consistent and asymptotically normal: As* $n \to \infty$*, we have* $\widehat{\zeta}_j \xrightarrow{p} \zeta_j$ *and*

$$\sqrt{n}(\widehat{\zeta}_j - \zeta_j) \xrightarrow{d} \mathcal{N}(0, V_{jj})$$

for all $j \in \mathcal{J}$*, where* V_{jj} *is the* j*-th diagonal element of the asymptotic variance-covariance matrix* $\boldsymbol{V}$ *defined in Appendix E.3.*

In addition, suppose that the margin condition in Assumption A.13 holds. Then, the estimator for the upper and lower bounds of the relative effect are consistent and asymptotically jointly normal: As $n \to \infty$*, we have* $\widehat{\tau}_L \xrightarrow{p} \tau_L$*,* $\widehat{\tau}_U \xrightarrow{p} \tau_U$*, and*

$$\sqrt{n}\begin{pmatrix} \widehat{\tau}_L - \tau_L \\ \widehat{\tau}_U - \tau_U \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} V_{LL} & V_{LU} \\ V_{UL} & V_{UU} \end{pmatrix}\right).$$

Proof is in Appendix E.3. The first part of the proposition establishes the consistency and asymptotic normality of the estimator for the distributional treatment effect ζ_j. The second part establishes the same properties for the estimators of the bounds on the relative effect τ. The results follow from a standard application of the M-estimation theory.

Statistical inference Based on the asymptotic normality established in Proposition 2, we can conduct statistical inference for the proposed estimators. For the distributional treatment effect ζ_j, we can construct the $(1-\alpha)$ confidence interval. The asymptotic variance-covariance matrix $\boldsymbol{V}$ can be consistently estimated using the nonparametric bootstrap method or by computing the sample analog of the variance formula derived in Appendix E.3.

For the relative effect τ, I construct the confidence interval for the partially identified parameter using the method proposed by Imbens and Manski (2004). Specifically, let $\widehat{\tau}_L$ and $\widehat{\tau}_U$ be the estimators for the lower and upper bounds of τ, respectively. The $(1-\alpha)$ confidence interval for τ is given as

$$C_{\tau,1-\alpha} = \left[\widehat{\tau}_L - z_{1-\alpha}\sqrt{\widehat{V}_{LL}/n}, \quad \widehat{\tau}_U + z_{1-\alpha}\sqrt{\widehat{V}_{UU}/n}\right]$$

where $\widehat{V}_{LL}$ and $\widehat{V}_{UU}$ are consistent estimators of V_{LL} and V_{UU}, respectively, and $z_{1-\alpha}$ is the critical value that solves $\Phi(z_{1-\alpha}+\delta) - \Phi(-z_{1-\alpha}) = 1-\alpha$ with $\delta = \sqrt{n}(\widehat{\tau}_U - \widehat{\tau}_L)/\max\{\sqrt{\widehat{V}_{LL}}, \sqrt{\widehat{V}_{UU}}\}$. This confidence interval covers the true parameter τ with at least $1-\alpha$ probability.

3 Assessing the distributional parallel trends assumption

This section develops a statistical procedure to assess the validity of the key identification assumption using pre-treatment periods. As in the standard DID design (e.g., Angrist and Pischke, 2008), additional pre-treatment periods provide an opportunity to assess the distributional parallel trends assumption (Assumption 3). Specifically, I propose an *equivalence-based testing* procedure to assess the plausibility of the distributional parallel trends assumption (Assumption 3). If the assumption holds in the pre-treatment periods, it is more reasonable to claim that it holds in the post-period. In Appendix B.5, I also develop a Wald-based test as an alternative approach to assess the assumption.

3.1 The proposed equivalence 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.

The goal is to assess if the distributional parallel trends assumption holds in the pre-treatment periods. Let $\widetilde{q}_d(v) = F_{d0}(F_{d1}^{-1}(v))$ denote the pre-treatment analog of $q_d(v)$ defined in Assumption 3. Recall that $q_d(v)$ captures the shift of distributions over time evaluated at quantile v, and the assumption requires $q_1(v) = q_0(v)$ for all $v \in (0,1)$. I propose an equivalence-based testing procedure to assess if $\widetilde{q}_1(v) = \widetilde{q}_0(v)$ holds for all v during the pre-treatment period. The intuition

is that if the maximum deviation between $\widetilde{q}_1$ and $\widetilde{q}_0$ is "small" enough, then we may conclude that the two functions are equivalent and the assumption is more plausible. Let $r(v) = \widetilde{q}_1(v) - \widetilde{q}_0(v)$ and $\|r\|_\infty = \sup_{v\in(0,1)} |r(v)|$. For some threshold value of $\delta > 0$, the equivalence test is formulated as the following hypotheses:

$$H_0 \colon \|r\|_\infty \geq \delta \quad \text{and} \quad H_1 \colon \|r\|_\infty < \delta \tag{3.1}$$

where H_0 says that the maximum deviation between two functions is larger than an acceptable level of equivalence δ, and H_1 says that the maximum deviation is smaller than δ.

The equivalence-based approach is appealing to test the condition, since we do not conflate the power of the test with the evidence for the assumption (see Hartman and Hidalgo, 2018). The test is constructed such that the null hypothesis implies that the distributional parallel trends assumption does not hold with an acceptable level of deviation δ. Therefore, rejecting the null provides direct evidence *in favor of the identification assumption* (or more precisely the pre-treatment analog of the assumption). When the test fails to reject the null, either because of the violation of the assumption or lack of power, we conclude that there is not enough evidence to support the assumption. A similar approach has been considered in the DID literature to test the parallel trends assumption (e.g., Egami and Yamauchi, 2023; Liu, Wang and Xu, 2024).

For now, we assume that a researcher has selected a value of δ to assess the equivalence. In the next section, I explore a practical approach to choose a reasonable value of δ. Specifically, I show that the bias of the treatment effects can be bounded as a function of δ, and that researchers can choose δ based on the acceptable level of bias in the estimated effects.

The above test in (3.1) can be cast as a standard two one-sided test (TOST) problem (Wellek, 2010). Specifically, the null H_0 is a union of two one-sided hypotheses: $H_0 = H_0^+ \cup H_0^-$ where

$$H_0^+ \colon \sup_{v\in(0,1)} r(v) \geq \delta \quad \text{and} \quad H_0^- \colon \inf_{v\in(0,1)} r(v) \leq -\delta.$$

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^-. For example, the null H_0^+ is rejected at level α, if the largest upper confidence interval is smaller than δ, that is,

$$\sup_v U_{1-\alpha}(v) < \delta$$

where $U_{1-\alpha}(v)$ is the upper bound of a $100(1-\alpha)\%$ point-wise confidence interval for $\widehat{r}(v)$.

The test statistics $\widehat{r}(v)$ can be estimated by estimating parameters μ_{dt} and σ_{dt} from the pre-treatment data. The estimating procedure is similar to that in Section 2.3, but I present the details

of the procedure in Appendix B.1 for completeness. The confidence intervals $U_{1-\alpha}(v)$ and $L_{1-\alpha}(v)$ can be computed by evaluating the variance formula in Lemma A.6 or by bootstrap.

Proposition 3 shows that the proposed procedure is in fact an asymptotically level α test. The test rejects the null of non-equivalence with probability less than α when the null is true (i.e., type I error is controlled) for any choice of δ that is consistent with the null.

Proposition 3 (Asymptotic size control of the test)**.** *Consider a TOST that rejects H_0 in (3.1) at level α iff $\mathcal{R}_+ \cap \mathcal{R}_-$ where $\mathcal{R}_+$ and $\mathcal{R}_-$ are the rejection events for H_0^+ and H_0^-, respectively:*

$$\mathcal{R}_+ = \left\{ \sup_{v\in(0,1)} U_{1-\alpha}(v) < \delta \right\} \quad \text{and} \quad \mathcal{R}_- = \left\{ \inf_{v\in(0,1)} L_{1-\alpha}(v) > -\delta \right\}.$$

Here, $U_{1-\alpha}(v)$ and $L_{1-\alpha}(v)$ are the upper and lower bounds of a $100(1-\alpha)\%$ point-wise confidence interval for $\widehat{r}(v)$. Then, for any δ that is consistent with the null hypothesis,

$$\sup_{P\in H_0} P\big(\mathcal{R}_+ \cap \mathcal{R}_-\big) \leq \alpha + o(1).$$

as $n \to \infty$.

In Proposition A.2, I show that the test is also consistent: the power of the test goes to one as the sample size increases when the alternative hypothesis is true (i.e., when the assumption is likely to hold). In Appendix B.2, I provide a power analysis of the proposed test, which shows that the power of the test is increasing in the distance between δ and the true maximum deviation $\|r\|_\infty$.

I can construct an equivalence confidence interval by inverting the above test. Specifically, the equivalence confidence interval is given by $[-\widehat{\delta}, \widehat{\delta}]$ where $\widehat{\delta}$ corresponds to the smallest value of δ such that we reject the null of non-equivalence at level α:

$$\widehat{\delta} = \max\left\{ \sup_{v\in(0,1)} U_{1-\alpha}(v), - \inf_{v\in(0,1)} L_{1-\alpha}(v) \right\}.$$

The value $\widehat{\delta}$ is informative in interpreting the test result. For example, if the estimated value is $\widehat{\delta} = 0.03$, then we can conclude any value of δ larger than 0.03 (e.g., $\delta = 0.1$) would lead to the rejection of the null of non-equivalence at level α (i.e., we practically conclude that the assumption is plausible given the specified value of δ). In Appendix B.3, I also provide a procedure to compute the p-value associated with the test.

3.2 Calibrating an equivalence threshold δ based on the worst case bias

So far, we have assumed that researchers have a clear idea what value should be used to assess the equivalence. In this section, I discuss how to choose a reasonable value of δ in practice. In

particular, I motivate the selection of δ based on the relationship between the value of δ and the potential bias in the estimated causal effects.

In Proposition A.3, I show that the bias of the estimated effects is bounded as a function of δ:

$$|\text{Bias}(\widehat{\zeta}_j)| \leq 2\delta/M + o_p(1) \quad \text{and} \quad |\text{Bias}(\widehat{\Delta}_j)| \leq \delta/M + o_p(1) \tag{3.2}$$

where $M = \inf_{v\in(0,1)} q_0'(v)$ is a constant that depends on the slope of $q_0(\cdot)$ and can be estimated from the data. In Appendix B.4, I provide a procedure to estimate M from the data.

The result in (3.2) provides a useful guideline to choose δ based on the acceptable level of bias in the estimated effects. For example, suppose that a researcher considers one percentage point as the acceptable level of bias in the estimated causal effects. Then, they can choose δ such that $2\delta/M = 0.01$ or equivalently $\delta = 0.005M$.

4 Extensions

4.1 Semiparametric Estimator with Covariates

When a rich set of pre-treatment covariates is available, we can relax the parametric assumption on the base distribution F in Assumption 2. Let $\boldsymbol{X}_i$ be a vector of observed pre-treatment covariates for unit i. Given the covariates, we can assume that the distributional parallel trends assumption holds conditionally on $\boldsymbol{X}_i$. Specifically, we can modify Assumption 3 as follows.

Assumption 4 (Conditional Distributional Parallel Trends)**.** For all $v \in (0,1)$ and $\boldsymbol{x} \in \mathcal{X}$,

$$q_1(v \mid \boldsymbol{x}) = q_0(v \mid \boldsymbol{x}),$$

where $q_d(v \mid \boldsymbol{x}) = F_{d0}(F_{d1}^{-1}(v \mid \boldsymbol{x}) \mid \boldsymbol{x})$ is the conditional quantile-quantile transform defined using the conditional distributions $F_{dt}(y \mid \boldsymbol{x}) = \Pr(Y_{it}^*(0) \leq y \mid D_i = d, \boldsymbol{X}_i = \boldsymbol{x})$.

A similar assumption that conditions on covariates has been discussed in Sofer et al. (2016) in the context of negative control outcomes. Here, I extend the assumption on the latent variable Y^* given observed covariates. The identification formula based on this assumption is presented in Appendix C.

To estimate the model with covariates, I propose the following specification of the parameters:

$$\mu_{dt}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_{dt}, \quad \text{and} \quad \sigma_{dt}(\boldsymbol{x}) = \exp(\boldsymbol{x}^\top \boldsymbol{\xi}_{dt}).$$

Under the location-scale model, modeling the mean and the standard deviation as functions of covariates is sufficient to characterize the entire conditional distribution. I estimate the base dis-

tribution F nonparametrically using the NPMLE under the interval-censoring model induced by the ordered response (Cosslett, 1983; Liu and Yu, 2024). Although the estimator depends on the nonparametric estimate of the base distribution $\widehat{F}$, I show that the treatment effect estimator $\widehat{\zeta}_j$ attains asymptotic normality at the parametric rate (Proposition A.5). The estimation procedure, asymptotic theory, and simulation results are presented in Appendix C.

4.2 Staggered Adoption

I extend the proposed approach to a staggered adoption setup. In many policy studies, the treatment is adopted at different times across units, making the staggered adoption design one of the most common data structures in applied research (e.g., Callaway and Sant'Anna, 2021).

Let $G_i \in \{1, \ldots, T, \infty\}$ denote the time period when unit i first receives the treatment, with $G_i = \infty$ for never-treated units. The identifying assumption requires that the quantile-quantile transform between any two time periods is identical for the treated group $G_i = g$ and the never-treated group $G_i = \infty$. This is a natural extension of the distributional parallel trends assumption (Assumption 3) to the staggered adoption setup, analogous to the group-specific parallel trends assumption discussed in Callaway and Sant'Anna (2021). The target estimands are group-time specific effects $\zeta_j(g,t)$ and $\tau(g,t)$, which measure the category-specific and relative effects for the group first treated at time g, evaluated at time $t \geq g$. These can be aggregated into overall effects $\overline{\zeta}_j$ and $\overline{\tau}$ using researcher-specified weights (see Callaway and Sant'Anna, 2021, for a discussion on aggregation). The formal setup, identification, and estimation procedure are presented in Appendix D.

5 Empirical Application

As discussed in the introduction, a recent debate examines the effect of mass shootings on public support for gun control regulations (Newman and Hartman, 2019; Barney and Schaffner, 2019; Hartman and Newman, 2019). The attitude toward gun control is measured via a survey question in the Cooperative Congressional Election Study (CCES) (Kuriwaki, 2018; Schaffner and Ansolabehere, 2015):

> 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.`

Throughout the debate, the authors utilize a variety of methodologies, such as an ordered probit model with random effects and a linear fixed effect model, to estimate the impact of mass shootings on the attitude (see Table 2 in Appendix G).

I first reanalyze the two-wave panel of CCES (2010–2012). The two-wave panel is the primary dataset for the panel-based analyses in the original debate: Newman and Hartman (2019) use it as their main panel analysis, and both Barney and Schaffner (2019) and Hartman and Newman (2019) center their replication and reanalysis on this dataset. I then analyze the three-wave panel of CCES (2010–2012–2014), which allows us to assess the identification assumption using the additional pre-treatment period.

I estimate the category-specific effects ζ_j for $j = 0, 1, 2$ and the relative effect τ as defined in Section 2. Recall that `less-strict` is coded as `0` and `more-strict` is coded as `2`. A positive τ indicates that mass shootings increase the likelihood of preferring stricter regulations.

5.1 Results from the two-wave panel

This section presents the results of the analysis on the two-wave sample from CCES ($n = 16,553$). The outcome is measured in 2010 and 2012 and I treat a response in 2012 as the post-treatment outcome. I consider two definitions of the treatment variable based on the distance from the mass shooting incidents. In the first definition, respondents living in a neighborhood where mass shootings happened within 25 miles between 2010 and 2012 are considered as treated ($n_1 = 1,611$). I also use 100 miles as an alternative threshold for defining the treatment group ($n_1 = 4,877$). This mirrors Barney and Schaffner (2019)'s approach who considered multiple distance thresholds to define the treatment group. In total, there were 16 mass shooting incidents recorded in the dataset between the two waves of CCES (Newman and Hartman, 2019, Appendix C).

I also investigate effect heterogeneity by pre-treatment covariates. First, following the original papers, I investigate if effects vary across respondents' party affiliations. Second, I investigate if the baseline safety of the neighborhood affects how people respond to mass shootings. Respondents are classified into either "prior exposure" or "no prior exposure" group based on whether mass shootings occurred within 100 miles of their area in the last ten years (as of 2010).

Figure 2 shows the main results. The left panel of the figure shows the estimated effect $\widehat{\zeta}_j$ for $j = 0, 1, 2$. The red circles represent estimates for the 100 mile threshold, while blue triangles represent estimates for the 25 mile threshold. The estimates $\widehat{\zeta}_0$ capture the effect on the probability of preferring less strict gun regulations, while $\widehat{\zeta}_2$ captures the effect on the probability of preferring more strict control of firearm sales. Along with point estimates, I also show the 95% confidence intervals, which are computed using $B = 5{,}000$ block bootstrap draws clustered at the zip code level.

The figure shows that in both definitions of the treatment coding, we see a decrease in the middle category (`keep-as-they-are`), while there are small increases in both tails (`less-strict` and `more-strict`). However, the confidence intervals cover zero for all estimates.

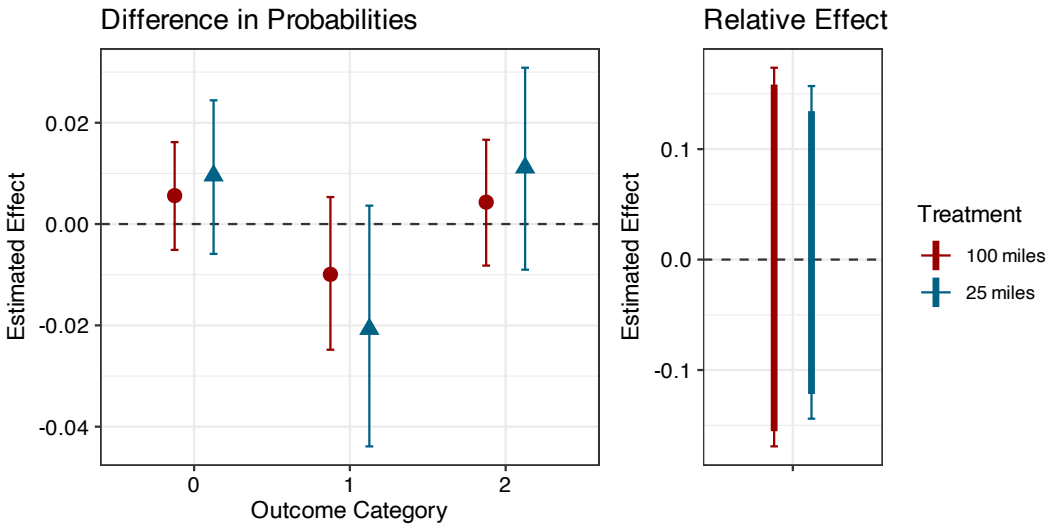

Figure 2: Left – Estimated treatment effects $\widehat{\zeta}_j$ with 95% confidence intervals. Red circles indicate effect for the 100 mile threshold, while blue triangles indicate effect for the 25 mile threshold. Labels above estimates indicate subsamples used for the analysis where n indicates the size of the sample. **Right** – Estimated bounds on the relative effect $[\widehat{\tau}_L, \widehat{\tau}_U]$ (solid lines) with 95% confidence intervals (thin lines). Red lines indicate effect for the 100 mile threshold, while blue lines indicate effect for the 25 mile threshold.

The right panel shows the estimated bounds on the relative effect $\widehat{\tau}$. The identification region always contains zero, so we cannot rule out no overall shift in the distribution of the outcome.

Overall, the results do not provide strong evidence that experiencing nearby mass shootings affects people's attitudes toward gun control regulations in either direction. The original studies reach different conclusions using ordered logit with random effects, ordered logit with fixed effects, and linear two-way fixed effects (see Table 2 in Appendix G), none of which directly targets the ordinal structure under a transparent identification assumption. In Appendix G.2, I also apply the semiparametric estimator from Section 4.1 with pre-treatment covariates.

Figure 3 shows the results of subgroup analysis by partisanship. The panels in the first row show the estimates for $\widehat{\zeta}_j$, where columns correspond to the three categories of the partisanship. The figure shows that effects for Democrats (left panel) and Republicans (right panel) are in the expected direction: Democrats tend to move toward stricter regulations while Republicans tend to move toward less strict regulations, and the pattern is more pronounced when we use the 25 mile threshold. However, across the estimates, the 95% confidence intervals always cover zero.

The panels in the second row of Figure 3 show the results for the relative effect $\widehat{\tau}$. Appendix G.3 reports a similar analysis stratified by prior exposure to mass shootings.

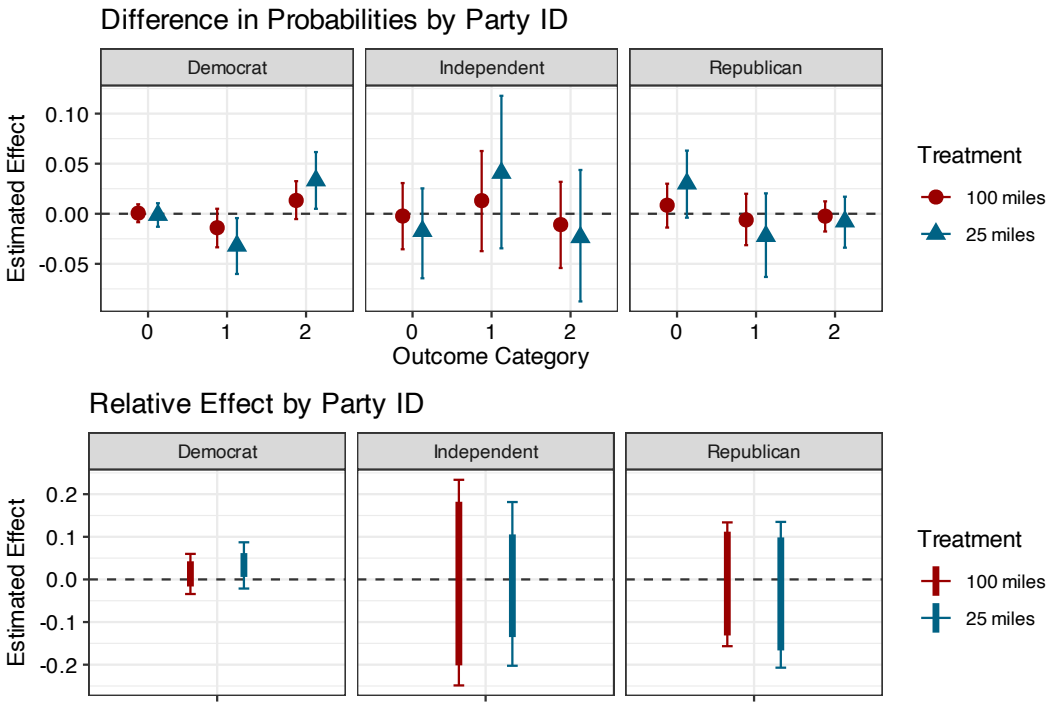

Figure 3: Upper panels – Estimated treatment effects $\widehat{\zeta}_j$ for $j = 0, 1, 2$ by party identification with 95% confidence intervals. Circles indicate effects under the 100 mile threshold, while triangles indicate effects under the 25 mile threshold. **Lower panel** – Estimated bounds on the relative effect $[\widehat{\tau}_L, \widehat{\tau}_U]$ by party identification (thick lines) with 95% confidence intervals (thin lines).

5.2 Diagnostics using three-wave panel

To demonstrate the practical use of the proposed diagnostic test, I analyze the three-wave panel from CCES (2010-12-14) in this section. The goal is to use the pre-treatment periods (2010 and 2012 waves) to assess if the identification assumption in Assumption 3 is plausible.

I create a pre-treatment sample from the three-wave panel data by including only two types of respondents: (1) those who never experience the mass shootings throughout the sample periods (control group) and (2) those who experienced the mass shootings only after 2012 (treated group). To avoid the possibility that the past exposure might affect the baseline attitudes, I further condition on respondents in the "no prior exposure" group. I report the details of the sample construction in Appendix G.

I apply the diagnostic test proposed in Section 3 to the pre-treatment outcome between 2010 and 2012. The goal here is to statistically test if $\widetilde{q}_1(v) = \widetilde{q}_0(v)$ holds for all $v \in (0, 1)$, where $\widetilde{q}_d(v)$ is the pre-treatment analog of the quantile-quantile relationship defined on group d. Specifically, I test the null hypothesis of non-equivalence, $H_0 \colon \|r\|_\infty \geq \delta$.

Figure 4 shows the estimated differences between the two quantile-quantile transformations $\widehat{r}(v)$ (solid line) with the point-wise 95% confidence intervals (dashed lines). The left panel shows the result for the 100 mile threshold, and the right panel is for the 25 mile threshold.

To facilitate the interpretation of the results, I also show the equivalence confidence intervals $[-\widehat{\delta}, \widehat{\delta}]$ (red lines) in the figure. The equivalence confidence interval corresponds to the rejection threshold δ such that any choice of δ larger than $\widehat{\delta}$ leads to the rejection of the null at the 5% level. For example, suppose that we pick $\delta = 0.1$ as our equivalence threshold. Then, in both definitions of the treatment, we can reject the null of non-equivalence at the 5% level since the equivalence confidence interval is strictly contained in $[-0.1, 0.1]$.

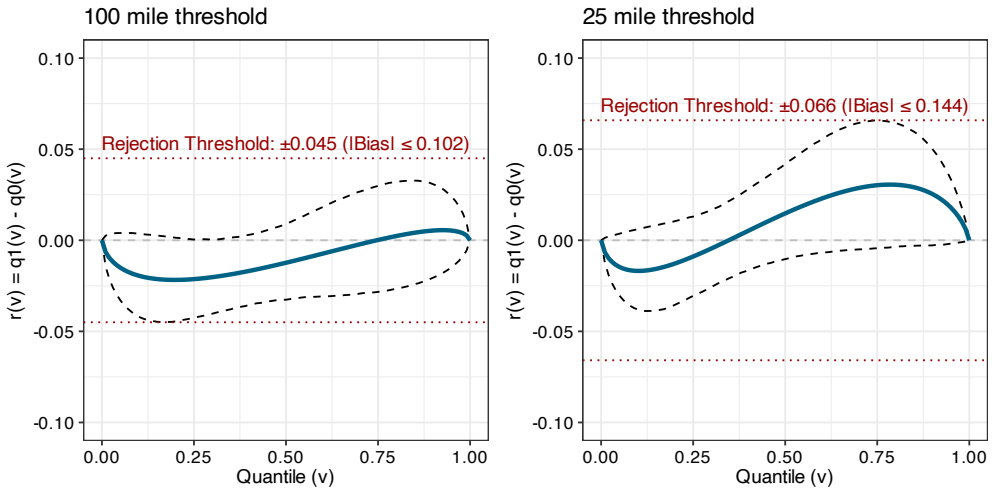

Figure 4: Estimated $\hat{r}(v)$ (solid line) with point-wise 95% confidence intervals (dashed lines). Red lines show the equivalence confidence intervals $[-\delta, \delta]$. The left panel shows the result for the 100 mile threshold, and the right panel is for the 25 mile threshold. The result suggests that the null is rejected at the 5% level with $\delta = 0.044$ for the 100 mile threshold and $\delta = 0.064$ for the 25 mile threshold.

The result suggests that the null of non-equivalence is rejected at the 5% level when we choose $\delta \geq 0.044$ for the 100 mile threshold and $\delta \geq 0.064$ for the 25 mile threshold. Applying the worst case bias formula in (3.2), these values of δ correspond to the maximum biases of approximately 10.1 and 14.1 percentage points. In other words, if a researcher is willing to tolerate these levels of biases in the estimated effects, then the identification assumption would be considered as plausible based on the pre-treatment data.

Taken together, the two-wave analysis finds no strong evidence of a uniform shift in either direction, and the three-wave diagnostic suggests that the identification assumption is plausible when the researcher tolerates moderate levels of bias.

6 Concluding Remarks

Despite recent progress in the DID literature, ordinal outcomes have received limited methodological attention. In this paper, I proposed a latent variable approach that formulates distributional shift on the continuous variable underlying the ordinal outcome and identifies the entire counterfactual distribution without imposing linearity on the observed outcome scale. I also developed an equivalence-based testing procedure that connects the test threshold to the worst-case bias in the treatment effect estimate, so that researchers can assess whether the identification assumption is plausible.

The proposed framework involves several trade-offs that practitioners should consider. The basic model assumes a parametric distribution for the latent variable, which could introduce bias when the distribution for the latent variable is misspecified. The semiparametric extension in Section 4 relaxes this assumption at the cost of requiring covariates and a larger sample size. The proposed method is most useful when researchers have ordinal outcomes with a moderate number of categories and panel data with at least three time periods (two pre-treatment periods and one post-treatment period). When the outcome is continuous, the distributional DID approach of Athey and Imbens (2006) or the existing methods based on the parallel trends assumptions can be directly applied.

Several directions for future research remain open. First, the current framework handles pre-treatment covariates, but extending it to time-varying confounders would let researchers handle settings where covariates evolve alongside the treatment. Second, extending the method for multiple ordinal outcomes would be useful for applications where the attitudes are measured through multiple questions. The latent variable framework proposed in this paper would be well-suited for this extension.

The open-source `R` package `orddid` implements the proposed methodology and is available at `https://github.com/soichiroy/orddid`.

References

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

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

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

Barney, David J and Brian F Schaffner. 2019. "Reexamining the Effect of Mass Shootings on Public Support for Gun Control." *British Journal of Political Science* 49(4):1555–1565.

Callaway, Brantly and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225(2):200–230.

Callaway, Brantly, Tong Li and Tatsushi Oka. 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.

Card, D and AB Krueger. 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.

Chalmers, R Philip. 2012. "`mirt`: A Multidimensional Item Response Theory Package for the `R` Environment." *Journal of Statistical Software* 48(6):1–29.

Chang, Neng-Chieh. 2020. "Double/debiased machine learning for difference-in-differences models." *The Econometrics Journal* 23(2):177–191.

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

Cosslett, Stephen R. 1983. "Distribution-free maximum likelihood estimator of the binary choice model." *Econometrica: Journal of the Econometric Society* pp. 765–782.

Egami, Naoki and Soichiro Yamauchi. 2023. "Using multiple pretreatment periods to improve difference-in-differences and staggered adoption designs." *Political Analysis* 31(2):195–212.

Geskus, Ronald and Piet Groeneboom. 1999. "Asymptotically optimal estimation of smooth functionals for interval censoring, case 2." *The Annals of Statistics* 27(2):627–674.

Glynn, Adam and Nahomi Ichino. 2019. "Generalized Nonlinear Difference-in-Difference-in-Differences." *Working Paper* .

Groeneboom, Piet and Kim Hendrickx. 2018. "Current status linear regression." *The Annals of Statistics* 46(4):1415–1444.

Hartman, Erin and F Daniel Hidalgo. 2018. "An equivalence approach to balance and placebo tests." *American Journal of Political Science* 62(4):1000–1013.

Hartman, Todd K and Benjamin J Newman. 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.

Imbens, Guido W and Charles F Manski. 2004. "Confidence intervals for partially identified parameters." *Econometrica* 72(6):1845–1857.

Kuriwaki, Shiro. 2018. "Cumulative CCES Common Content (2006-2018).".
URL: *https://doi.org/10.7910/DVN/II2DB6*

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

Liu, Licheng, Ye Wang and Yiqing Xu. 2024. "A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data." *American Journal of Political Science* 68(1):160–176.

Liu, Ruixuan and Zhengfei Yu. 2024. "Simple Semiparametric Estimation Of Ordered Response Models." *Econometric Theory* 40(1):1–36.

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

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

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

Lu, Jiannan, Yunshu Zhang and Peng Ding. 2020. "Sharp bounds on the relative treatment effect for ordinal outcomes." *Biometrics* 76(2):664–669.

Maddala, Gangadharrao Soundalyarao. 1983. "Limited-dependent and qualitative variables in econometrics." *Econometric Society Monographs* 3.

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

Park, Chan and Eric Tchetgen Tchetgen. 2022. "A universal difference-in-differences approach for causal inference." *arXiv preprint arXiv:2212.13641* .

Richardson, David B, Ting Ye and Eric J Tchetgen Tchetgen. 2023. "Generalized difference-in-differences." *Epidemiology* 34(2):167–174.

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

Samejima, Fumiko. 1969. *Estimation of Latent Ability Using a Response Pattern of Graded Scores.* Psychometrika. Psychometrika Monograph Supplement No. 17.

Schaffner, Brian and Stephen Ansolabehere. 2015. "2010-2014 Cooperative Congressional Election Study Panel Survey.".
URL: *https://doi.org/10.7910/DVN/TOE8I1*

Sofer, Tamar, David B Richardson, Elena Colicino, Joel Schwartz and Eric J Tchetgen Tchetgen. 2016. "On negative outcome control of unobserved confounding as a generalization of difference-in-differences." *Statistical Science* 31(3):348.

Tchetgen Tchetgen, Eric J, Chan Park and David B Richardson. 2024. "Universal difference-in-differences for causal inference in epidemiology." *Epidemiology* 35(1):16–22.

Train, Kenneth E. 2009. *Discrete choice methods with simulation.* Cambridge University Press.

Turnbull, Bruce W. 1976. "The empirical distribution function with arbitrarily grouped, censored and truncated data." *Journal of the Royal Statistical Society: Series B (Methodological)* 38(3):290–295.

van der Vaart, AW. 2000. Asymptotic Statistics. Technical report Cambridge University Press.

Volfovsky, Alexander, Edoardo M Airoldi and Donald B Rubin. 2015. "Causal inference for ordinal outcomes." *arXiv preprint arXiv:1501.01234* .

Wellek, Stefan. 2010. *Testing statistical hypotheses of equivalence and noninferiority.* Chapman and Hall/CRC.

Wellner, Jon A and Yihui Zhan. 1997. "A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data." *Journal of the American Statistical Association* 92(439):945–959.

Appendix

A Causal Estimands for Ordinal Outcomes

A.1 Sharp bound on the relative effect

Following the notation in Lu, Zhang and Ding (2020), write the marginal distribution of the potential outcomes for the treated group as

$$p_{k+} \equiv \Pr(Y_{i1}(1) = k \mid D = 1) \qquad p_{+l} \equiv \Pr(Y_{i1}(0) = l \mid D = 1)$$

for $k, l \in \{0, \ldots, J-1\}$. The following proposition presents the sharp bounds on the relative effect τ. The only difference from Lu, Zhang and Ding (2020) is that I consider the treatment effect on the treated group, thus replacing all the marginal probabilities with those conditional on $D = 1$.

Proposition A.1 (Theorem 1 of Lu, Zhang and Ding, 2020)**.** *For $j = 1, \ldots, J-1$ and $m = 1, \ldots, J-j$, define*

$$\delta_{jm} = \sum_{k=j}^{J-1} p_{k+} + \sum_{k=j+m}^{J-1} p_{k+} + \sum_{l=0}^{j-2} p_{+l} - \sum_{l=j+m-1}^{J-1} p_{+l},$$

with the convention that a sum over an empty index set equals 0. Then the sharp upper bound is

$$\tau_U = \min_{1 \le j \le J-1} \min_{1 \le m \le J-j} \delta_{jm}.$$

Define, for the same index ranges,

$$\begin{aligned} \xi_{jm} &= -\left(\sum_{l=j}^{J-1} p_{+l} + \sum_{l=j+m}^{J-1} p_{+l} + \sum_{k=0}^{j-2} p_{k+} - \sum_{k=j+m-1}^{J-1} p_{k+} \right) \\ &= \sum_{k=j+m-1}^{J-1} p_{k+} - \sum_{k=0}^{j-2} p_{k+} - \sum_{l=j}^{J-1} p_{+l} - \sum_{l=j+m}^{J-1} p_{+l}. \end{aligned}$$

Then the sharp lower bound is given by

$$\tau_L = \max_{1 \le j \le J-1} \max_{1 \le m \le J-j} \xi_{jm}.$$

We can express the upper and lower bounds in a more compact form. Let $D_U(j) = \Pr(Y_{i1}(1) \ge j \mid D = 1) - \Pr(Y_{i1}(0) \ge j \mid D = 1)$. Then,

$$\tau_U = 1 + \min_{1 \le j \le J-1} \min_{1 \le j' \le J-j} \{D_U(j) + D_U(j + j')\}$$

B Additional Results on Equivalence Testing

B.1 Implementation of the equivalence test

The proposed equivalence test consists of evaluating two rejection events:

$$\mathcal{R}_+ = \left\{ \sup_{v\in(0,1)} U_{1-\alpha}(v) < \delta \right\} \quad \text{and} \quad \mathcal{R}_- = \left\{ \inf_{v\in(0,1)} L_{1-\alpha}(v) > -\delta \right\}.$$

Evaluating these events requires constructing point-wise confidence intervals for $\widehat{r}(v)$.

This section discusses how to construct the confidence intervals from the data on the pre-treatment periods. Under Assumptions 1 and 2, we can express $\widetilde{q}_d(v)$ as a function of the model parameters $\boldsymbol{\theta}$:

$$\widetilde{q}_d(v) = F_U\left(\frac{\mu_{d1}-\mu_{d0}}{\sigma_{d0}} + \frac{\sigma_{d1}}{\sigma_{d0}} F_U^{-1}(v)\right)$$

for $d \in \{0,1\}$. Therefore, the test statistics $\widehat{r}(v)$ can be computed once the model parameters are estimated from the data.

Following the estimation procedure similar to Section 2.3, I estimate parameters as follows:

Step 1 Estimate μ_{00} and $\boldsymbol{\kappa}$ using the data from the control group in the pre-treatment period. I normalize $\sigma_{00} = 1$ and $\kappa_1 = 0$ for identification.

Step 2 Estimate $\boldsymbol{\theta}_{dt} = (\mu_{dt}, \sigma_{dt})$ for $(d,t) \in \{(0,1),(1,0),(1,1)\}$ by maximizing the likelihood

$$\widehat{\boldsymbol{\theta}}_{dt} = \arg\max_{\mu_{dt},\sigma_{dt}} \sum_{i:D_i=d} \sum_{j\in\mathcal{J}} \mathbf{1}\{Y_{it}=j\} \log\left\{ F((\widehat{\kappa}_{j+1}-\mu_{dt})/\sigma_{dt}) - F((\widehat{\kappa}_j-\mu_{dt})/\sigma_{dt}) \right\}$$

where I use the estimated cutoffs $\widehat{\boldsymbol{\kappa}}$ from Step 1.

Step 3 Compute $\widehat{r}(v)$ on the grid of $v \in (0,1)$ as

$$\widehat{r}(v) = \widehat{\widetilde{q}}_1(v) - \widehat{\widetilde{q}}_0(v)$$

where

$$\widehat{\widetilde{q}}_d(v) = F_U\left(\frac{\widehat{\mu}_{d1}-\widehat{\mu}_{d0}}{\widehat{\sigma}_{d0}} + \frac{\widehat{\sigma}_{d1}}{\widehat{\sigma}_{d0}} F_U^{-1}(v)\right).$$

Step 4 Construct a point-wise confidence interval for $\widehat{r}(v)$

$$U_{1-\alpha}(v) = \widehat{r}(v) + z_{1-\alpha}\sqrt{\widehat{\mathrm{Var}}(\widehat{r}(v))/n}$$

where the variance can be estimated using the variance formula in Lemma A.6 or the bootstrap. The lower bound $L_{1-\alpha}(v)$ is constructed similarly.

Lemma A.6 in Appendix E.1 establishes the validity of the above confidence interval construction by showing the point-wise asymptotic normality of $\widehat{r}(v)$.

Given the confidence intervals, the test can be conducted by following the procedure in Section 3. Furthermore, related statistics such as equivalence confidence intervals and p-values can be computed with the confidence intervals.

B.2 Consistency and power of the equivalence test

In Proposition A.2, I show that the proposed equivalence test is consistent.

Proposition A.2 (Consistency of the test)**.** *For fixed values of* $\alpha \in (0,1)$ *and* $\delta \in (0,1)$*, suppose that there exists* $\epsilon > 0$ *such that*

$$\|r(v)\|_\infty \leq \delta - \epsilon,$$

then, the TOST defined in Proposition 3 is consistent.

$$P(\mathcal{R}_+ \cap \mathcal{R}_-) \to 1$$

as $n \to \infty$*.*

Proof is in Section E.5. The result shows that when the null hypothesis under a fixed $\delta > 0$ is false (i.e., the distributional parallel trends assumption holds with some margin δ), the test will reject the null with probability approaching one. In other words, asymptotically, the test has power one against fixed alternatives.

Based on the above results, we can also express the asymptotic power function. Power is the probability of correctly rejecting the null hypothesis when the alternative is true. Let $v^+ = \arg\max r(v)$ and $v^- = \arg\min r(v)$ be the unique maximizer and minimizer of $r(v)$. Then, the power function in our context can be expressed as follows.

$$\begin{aligned}
\beta &= \Pr(\underbrace{U_{1-\alpha}(v^+) < \delta, L_{1-\alpha}(v^-) > -\delta}_{\text{rejection event}} \mid \underbrace{\|r(v)\|_\infty \leq \delta}_{\text{alternative true}}) \\
&= \Pr\left(\underbrace{\frac{\widehat{r}(v^+) - r(v^+)}{\widehat{\sigma}(v^+)/\sqrt{n}}}_{\equiv Z_+} + \frac{r(v^+) - \delta}{\widehat{\sigma}(v^+)/\sqrt{n}} < -z_{1-\alpha}, \underbrace{\frac{\widehat{r}(v^-) - r(v^-)}{\widehat{\sigma}(v^-)/\sqrt{n}}}_{\equiv Z_-} + \frac{r(v^-) + \delta}{\widehat{\sigma}(v^-)/\sqrt{n}} > z_{1-\alpha} \right) \\
&= \Pr\Big(Z_+ \leq a_+(\delta), -Z_- \leq a_-(\delta) \Big) \\
&= \Phi_2(a_+(\delta), a_-(\delta); -\rho)
\end{aligned}$$

where in the second equality I used the fact that when $\|r(v)\|_\infty \leq \delta$, it can be either $r(v^+) < \delta$ or $r(v^-) > -\delta$, with

$$a_+(\delta) = \frac{\delta - r(v^+)}{\widehat{\sigma}(v^+)/\sqrt{n}} - z_{1-\alpha}, \quad a_-(\delta) = \frac{\delta + r(v^-)}{\widehat{\sigma}(v^-)/\sqrt{n}} - z_{1-\alpha}.$$

and $\rho = \text{Corr}(\widehat{r}(v^+), \widehat{r}(v^-))$.

Intuitively, the power increases as the distance between $r(v^+)$ and δ, or $r(v^-)$ and $-\delta$, increases. In other words, when we specify a lenient threshold δ that is far away from the true maximum deviation $\|r(v)\|_\infty$, the test has higher power to reject the null hypothesis.

Figure B.1 shows the power curves under different sample sizes and correlation between $\widehat{r}(v^+)$ and $\widehat{r}(v^-)$. The figure shows a scenario with $\sigma(v^+) = \sigma(v^-) = 1$ and $r(v^+) = 0.2$, $r(v^-) = -0.15$. The x-axis is the distance between δ and $\|r(v)\|_\infty$, and the y-axis is the power.

As we expect, the figure shows that the power increases as the distance between δ and $\|r(v)\|_\infty$ increases. This means that when the sample size is small, we need to specify a lenient threshold δ to have sufficient power. Also, the power is lower when the correlation ρ is higher.

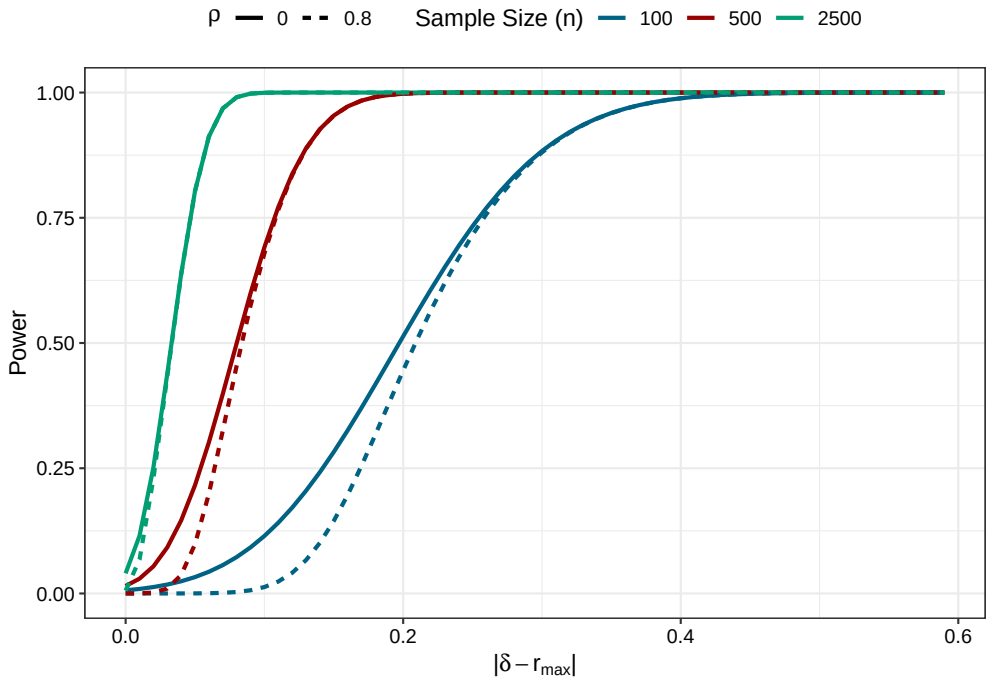

Figure B.1: Power curves of the equivalence test under different sample sizes and correlation between $\widehat{r}(v^+)$ and $\widehat{r}(v^-)$. The power is increasing in the distance between δ and $\|r(v)\|_\infty$, and in the sample size. The power is lower when the correlation is higher.

B.3 p-values for the equivalence test

The result in Proposition A.2 indicates that we can compute the p-value for the test by solving the rejection rule with respect to α,

$$\widehat{p} = \max\left\{ \max_{v\in[\epsilon,1-\epsilon]} \widehat{p}_1(v), \max_{v\in[\epsilon,1-\epsilon]} \widehat{p}_2(v) \right\},$$

where $[\epsilon, 1-\epsilon]$ is a grid over $(0,1)$ for small value of $\epsilon > 0$ chosen for numerical stability, where $\widehat{p}_1(v)$ and $\widehat{p}_2(v)$ are the point-wise p-values for testing:

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

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

B.4 Calibrating an equivalence threshold δ

I show that the bias of the estimated effects can be bounded as a function of the equivalence threshold δ.

Proposition A.3 (Worst case bias bound)**.** *Let* $\delta = \|r(v)\|_\infty$ *where* $r(v) = q_1(v) - q_0(v)$*. Then,*

the worst case bias of the estimated effects is bounded as

$$|Bias(\widehat{\zeta}_j)| \leq 2\delta/M + o_p(1) \quad \textit{and} \quad |Bias(\widehat{\Delta}_j)| \leq \delta/M + o_p(1)$$

where $M = \inf_{v\in(0,1)} q_0'(v)$.

The proof is in Section E.6.

Estimating M from the data To estimate $M = \inf_{v\in(0,1)} q_0'(v)$ from the data, I first derive the expression of $q_0'(v)$ as

$$\begin{aligned}\frac{d}{dv}q_0(v) &= f_U(a + bF_U^{-1}(v)) \cdot b \cdot \frac{d}{dv}F_U^{-1}(v) \\ &= b \cdot \frac{f_U(a + bF_U^{-1}(v))}{f_U(F_U^{-1}(v))}\end{aligned}$$

where $a = (\mu_{01} - \mu_{00})/\sigma_{00}$ and $b = \sigma_{01}/\sigma_{00}$.

Then, I plug in the MLE $\widehat{\boldsymbol{\theta}}_0 = (\widehat{\mu}_{00}, \widehat{\mu}_{01}, \widehat{\sigma}_{01})$, where $\sigma_{00} = 1$ is fixed, to obtain the estimator of $q_0'(v)$ as

$$\widehat{q}_0'(v) = \widehat{\sigma}_{01} \cdot \frac{\phi(\widehat{\mu}_{01} - \widehat{\mu}_{00} + \widehat{\sigma}_{01}\Phi^{-1}(v))}{\phi(\Phi^{-1}(v))}$$

where I take the base distribution $U \sim \mathcal{N}(0,1)$. Finally, I numerically solve $\min_{v\in[\epsilon,1-\epsilon]} \widehat{q}_0'(v)$ to obtain the estimate of M, where ϵ is a small positive constant to avoid the boundary issues.

B.5 Wald-based test for assessing the distributional parallel-trends assumption

Under Assumption 2, the distributional parallel trends assumption (Assumption 3) reduces to the restriction on the finite dimensional parameters. Specifically, $\widetilde{q}_d(v)$ can be written as

$$\widetilde{q}_d(v) = F_U\left(\frac{\mu_{d1} - \mu_{d0}}{\sigma_{d0}} + \frac{\sigma_{d1}}{\sigma_{d0}}F_U^{-1}(v)\right)$$

which implies that

$$\frac{\mu_{11} - \mu_{10}}{\sigma_{10}} = \frac{\mu_{01} - \mu_{00}}{\sigma_{00}} \quad \text{and} \quad \frac{\sigma_{11}}{\sigma_{10}} = \frac{\sigma_{01}}{\sigma_{00}} \implies \widetilde{q}_1(v) - \widetilde{q}_0(v) = 0.$$

Based on this observation, I consider the following Wald-type test:

$$H_0 \colon \psi(\boldsymbol{\theta}) = 0, \quad \text{vs} \quad H_1 \colon \psi(\boldsymbol{\theta}) \neq 0$$

where

$$\psi(\boldsymbol{\theta}) = \begin{bmatrix} (\mu_{11} - \mu_{10})/\sigma_{10} - (\mu_{01} - \mu_{00})/\sigma_{00} \\ \sigma_{11}/\sigma_{10} - \sigma_{01}/\sigma_{00} \end{bmatrix}.$$

The test statistic is given by

$$W = n \cdot \psi(\widehat{\boldsymbol{\theta}})^\top \left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\psi(\widehat{\boldsymbol{\theta}})\right)\widehat{\Omega}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\psi(\widehat{\boldsymbol{\theta}})\right)^\top\right]^{-1}\psi(\widehat{\boldsymbol{\theta}})$$

and reject the null when $W > \chi^2_{2,1-\alpha}$ where $\chi^2_{2,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-squared distribution with two degrees of freedom.

The asymptotic validity of the test (size control and power) follows from the standard Wald test argument and therefore the proof is omitted.

The Wald-based test can be used as an alternative approach to assess the distributional parallel trends assumption, and it is expected to be more powerful against the sup-norm test on $H_0 : \|r(v)\|_\infty = 0$. However, compared with the equivalence test proposed in Section 3, converting the test into an equivalence framework is not straightforward. For example, compared with the sup-norm test, the Wald statistic W does not have a direct interpretation in terms of bias (see the discussion in Section 3.2). Another option is to consider parametrizing the violation on the relationship between parameters:

$$H_0 \colon \begin{bmatrix} |(\mu_{11}-\mu_{10})/\sigma_{10} - (\mu_{01}-\mu_{00})/\sigma_{00}| \\ |\sigma_{11}/\sigma_{10} - \sigma_{01}/\sigma_{00}| \end{bmatrix} \geq \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$$

which requires specifying two equivalence thresholds $\delta_1, \delta_2 > 0$. Therefore, I leave the detailed investigation of the equivalence test based on the Wald statistic for future research.

B.6 Equivalence test with covariates

The equivalence testing procedure discussed in Section 3 can be extended to incorporate covariates. Specifically, we can assess the conditional distributional parallel trends assumption (Assumption 4) by testing if $\widetilde{q}_1(v \mid \boldsymbol{x}) = \widetilde{q}_0(v \mid \boldsymbol{x})$ holds for all $v \in [0,1]$ and $\boldsymbol{x} \in \mathcal{X}$ using pre-treatment periods.

Since it is not practical to specify the equivalence margin δ for all $\boldsymbol{x}$, I propose to focus on the average difference between two functions over the covariate distribution:

$$\overline{r} = \mathbb{E}[w(\boldsymbol{X}_i) r(\boldsymbol{X}_i)]$$

where $r(\boldsymbol{x}) = \sup_{v \in (0,1)} |\widetilde{q}_1(v \mid \boldsymbol{x}) - \widetilde{q}_0(v \mid \boldsymbol{x})|$. The weight function $w(\boldsymbol{x})$ is a user-specified function that satisfies $\mathbb{E}[w(\boldsymbol{X}_i)] = 1$. For example $w(\boldsymbol{x}) \propto f_{\boldsymbol{X}|D=1}(\boldsymbol{x})$ corresponds to the density of $\boldsymbol{x}$ in the treatment group. This implies that we weight the difference more heavily for covariate values that are more common in the treatment group.

The testing procedure follows as described in Section 3 where we test the null hypothesis $H_0 \colon \overline{r} \geq \delta$.

C Details on Semiparametric Approach with Covariates

C.1 Setup and identification

To include covariates $\boldsymbol{X}$ in the model, I modify the location-scale family assumption as

$$Y^*_{dt} \mid \boldsymbol{X} = \boldsymbol{x} \;\sim\; \mu_{dt}(\boldsymbol{x}) + \sigma_{dt}(\boldsymbol{x}) U, \quad U \sim F_U. \tag{C.1}$$

In addition, I make the following assumptions, which are standard in the semiparametric discrete choice literature.

Assumption A.5 (Sampling and panel structure)**.** $\{(Y_{i0}, Y_{i1}, D_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d. across i, and within-unit dependence across t is unrestricted. The proportion of the treated units satisfies $n_1/n \to p_1 \in (0,1)$.

Assumption A.6 (Ordered index and location-scale latent structure)**.** The latent index structure in Assumption 1 holds with covariates, and the latent utilities satisfy the location-scale family assumption as in (C.1) with $\mu_{dt}(\cdot)$ and $\sigma_{dt}(\cdot)$ known up to finite-dimensional parameters. Furthermore, the normalization is imposed as $\kappa_1 = 0$, $\sigma_{00} = 1$, and the first element of the coefficients $\beta_{00,1} = 1$.

Assumption A.7 (Conditions on the base distribution)**.** (i) The distribution F_0 is differentiable. (ii) F_0 is not constant over the support of $\boldsymbol{X}^\top \boldsymbol{\beta}_{00}$, and over the support of $\boldsymbol{X}^\top \boldsymbol{\beta}_{00} + \kappa_1$.

Assumption A.8 (Conditions on regressor and latent index)**.** Conditions ensuring identification of the first-stage single-index and F_0 hold:

(i) $\boldsymbol{X}$ and U are independent;

(ii) The first regressor X_1 has a Lebesgue density that is everywhere positive conditional on $\boldsymbol{X}_{-1}$;

(iii) Discrete components of $\boldsymbol{X}_{-1}$ satisfy overlap enabling support of $\boldsymbol{X}^\top \boldsymbol{\beta}$ to contain an open interval for each admissible $\boldsymbol{\beta}_{-1}$; and

(iv) The support of $\boldsymbol{X}^\top \boldsymbol{\beta}_{00}$ contains the support of U.

Proposition A.4 (Identification of parameters on Y_{00} and F)**.** *Under Assumptions A.6, A.7, and A.8, the parameters $(\boldsymbol{\beta}_{00}, \boldsymbol{\kappa})$ and the distribution F_U are identified.*

Proof. The result is a direct application of Liu and Yu (2024)'s Theorem 2.1. □

The following corollary extends Proposition 1 to the case with covariates that identifies the counterfactual distribution.

Corollary A.1 (Identification with covariates)**.** *Under Assumption 4 with conditions in Proposition A.4, the conditional counterfactual distribution of Y_{11}^* given $\boldsymbol{X} = \boldsymbol{x}$ is identified as*

$$\Pr(Y_{11} \le j \mid \boldsymbol{X} = \boldsymbol{x}) = F_U\left(\frac{\kappa_{j+1} - \mu_{11}(\boldsymbol{x})}{\sigma_{11}(\boldsymbol{x})}\right)$$

where $\mu_{11}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_{10} + \exp(\boldsymbol{x}^\top \boldsymbol{\xi}_{10}) \cdot \boldsymbol{x}^\top (\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{00})$, and $\sigma_{11}(\boldsymbol{x}) = \exp(\boldsymbol{x}^\top (\boldsymbol{\xi}_{10} + \boldsymbol{\xi}_{01}))$.

C.2 Semiparametric estimation

In the estimation and asymptotic analysis below, we write $F_0 \equiv F_U$ for the true base distribution, following standard semiparametric notation. The density is denoted f_0.

Let $\boldsymbol{X}_i = (X_{i1}, \boldsymbol{X}_{i,-1})$ denote the pre-treatment covariates where X_{i1} is a continuous variable. Similarly, let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{-1})$ denote the coefficient vector where β_1 is the coefficient for X_{i1}. In the following, I adopt the normalization $\beta_1 = 1$ to ensure the identification of the parameters, following the approach in Liu and Yu (2024). Following Liu and Yu (2024), the asymptotic theory below assumes $J = 3$ (i.e., three outcome categories). The extension to general J requires asymptotic theory for the interval-censoring NPMLE with more than two inspection times, which is not currently available (see Liu and Yu, 2024, Section S.3).

The estimation proceeds in two stages. The first stage recovers $(\boldsymbol{\beta}_{00}, \boldsymbol{\kappa})$ and the base distribution F from the pre-treatment control observations. The second stage estimates the remaining location-scale parameters $(\boldsymbol{\beta}_{01}, \boldsymbol{\xi}_{01})$ and $(\boldsymbol{\beta}_{10}, \boldsymbol{\xi}_{10})$ by profile likelihood conditional on $\widehat{F}$.

First stage: $(\boldsymbol{\beta}_{00}, \boldsymbol{\kappa})$ and F. An initial estimate $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_{00}$ is obtained by fitting an ordered probit model to $(Y_{i0}, \boldsymbol{X}_i)$ for $D_i = 0$. This also provides a preliminary cutoff estimate $\widetilde{\kappa}_2$. Given these values, the base distribution F is estimated as the nonparametric maximum likelihood estimator (NPMLE) under the interval-censoring model induced by the ordered response. Under the location-scale model with $\sigma_{00} = 1$ and $\kappa_1 = 0$, the observation $Y_{i0} = j$ for control unit i implies that the latent error $U_i = Y_{i0}^* - \boldsymbol{X}_i^\top \boldsymbol{\beta}_{00}$ lies in the interval $(L_i, R_i]$ where

$$L_i = \kappa_j - \boldsymbol{X}_i^\top \boldsymbol{\beta}_{00}, \quad R_i = \kappa_{j+1} - \boldsymbol{X}_i^\top \boldsymbol{\beta}_{00},$$

with the convention $\kappa_0 = -\infty$ and $\kappa_J = \infty$. For the three-category case ($J = 3$) with $\kappa_1 = 0$, we have

$$Y_{i0} = j \quad \Longrightarrow \quad U_i \in \begin{cases} (-\infty,\ -\boldsymbol{X}_i^\top \boldsymbol{\beta}_{00}] & \text{if } j = 0, \\ (-\boldsymbol{X}_i^\top \boldsymbol{\beta}_{00},\ \kappa_2 - \boldsymbol{X}_i^\top \boldsymbol{\beta}_{00}] & \text{if } j = 1, \\ (\kappa_2 - \boldsymbol{X}_i^\top \boldsymbol{\beta}_{00},\ \infty) & \text{if } j = 2. \end{cases} \tag{C.2}$$

The NPMLE of F under this interval-censoring model is

$$\widehat{F}_n(\cdot;\, \boldsymbol{\beta}, \boldsymbol{\kappa}) = \arg\max_{F \in \mathcal{F}} \sum_{i: D_i = 0} \log \left[F(R_i) - F(L_i) \right]$$

where $\mathcal{F}$ is the set of all CDFs on the real line, (L_i, R_i) are defined as in (C.2) evaluated at $(\boldsymbol{\beta}, \boldsymbol{\kappa})$, and the convention $F(-\infty) = 0$, $F(\infty) = 1$ is used for left- and right-censored observations. The NPMLE can be computed by the algorithms in Wellner and Zhan (1997) and Turnbull (1976).

This formulation uses all J outcome categories to estimate F, in contrast to the binary approach in Groeneboom and Hendrickx (2018) that splits $Y_{i0} = 0$ versus $Y_{i0} > 0$ and recovers F via isotonic regression on a single threshold. In simulations, I find that the NPMLE achieves substantially lower finite-sample bias in the estimated F.

Given $\widehat{F}_n$, the slope coefficients $\boldsymbol{\beta}_{00}$ are updated by solving the estimating equation

$$\Gamma(\boldsymbol{\beta}) = \frac{1}{n_0} \sum_{i: D_i = 0} \boldsymbol{X}_{i,-1} \Big\{ \mathbf{1}\{Y_{i0} = 0\} - \widehat{F}_n(-\boldsymbol{X}_i^\top \boldsymbol{\beta}) \Big\} = \mathbf{0}$$

and the cutoff κ_2 is then updated by solving

$$\Psi(\kappa_2) = \frac{1}{n_0} \sum_{i: D_i = 0} \Big\{ \mathbf{1}\{Y_{i0} \le 1\} - \widehat{F}_n(\kappa_2 - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_{00}) \Big\} = 0.$$

The NPMLE, the slope update, and the cutoff update are iterated until convergence. In practice, two iterations suffice.

Second stage: $(\boldsymbol{\beta}_{01}, \boldsymbol{\xi}_{01})$, $(\boldsymbol{\beta}_{10}, \boldsymbol{\xi}_{10})$, and the counterfactual distribution. I find that with the presence of the scale parameters $\boldsymbol{\xi}$, solving the estimating equations with the estimated CDF $\widehat{F}_n$ leads to substantial finite-sample bias in the second-stage parameters. Therefore, I propose to

estiamte the second-stage parameters by maximing a profile likelihood conditional on a regularized version of $\widehat{F}_n$. The NPMLE yields evaluation points $\{u_g, \widehat{F}_g\}_{g=1}^G$. Let $\underline{u} = \min_g u_g$ and $\overline{u} = \max_g u_g$. I define the regularized CDF by adding a vanishing linear ramp:

$$\widetilde{F}(u) = (1 - 2\varepsilon_n)\,\widehat{F}_n(u) + \varepsilon_n + \varepsilon_n \frac{u - \underline{u}}{\overline{u} - \underline{u}}, \qquad \varepsilon_n = n^{-1}. \tag{C.3}$$

This construction ensures that $\widetilde{F}$ is strictly increasing, bounded away from zero, and uniformly close to $\widehat{F}_n$ (Lemma A.10). In the implementation, I also replace the NPMLE estimates with I-spline smoothing. Specifically, the smoothed CDF is

$$\widetilde{F}^{\text{sp}}(u) = \widehat{F}(\underline{u}) + \sum_{k=1}^{K} \widehat{c}_k\, I_k(u)$$

where I_k are I-spline (integrated M-spline) basis functions of degree $d = 3$ with boundary knots at $\underline{u}$ and $\overline{u}$, and $K \propto n_0^{1/3}$ internal knots are placed at quantiles of the NPMLE grid. The coefficients $\widehat{\boldsymbol{c}}$ are obtained by constrained least squares with $c_k \geq 0$ for all k; the non-negativity constraint preserves monotonicity. The simulation in Section C.4 compares both regularizations. I find that the I-spline smoothing produces smaller finite sample biases.

Using the smoothed version of the estimated CDF $\widetilde{F}$, I maximize the profile likelihood to estimate $(\boldsymbol{\beta}_{01}, \boldsymbol{\xi}_{01})$ and $(\boldsymbol{\beta}_{10}, \boldsymbol{\xi}_{10})$:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\xi}; \widetilde{F}, \widehat{\boldsymbol{\kappa}}) = \prod_i \prod_{j=0}^{J-1} \left\{ \widetilde{F}\left(\frac{\widehat{\kappa}_{j+1} - \boldsymbol{X}_i^\top \boldsymbol{\beta}}{\exp(\boldsymbol{X}_i^\top \boldsymbol{\xi})} \right) - \widetilde{F}\left(\frac{\widehat{\kappa}_j - \boldsymbol{X}_i^\top \boldsymbol{\beta}}{\exp(\boldsymbol{X}_i^\top \boldsymbol{\xi})} \right) \right\}^{\mathbf{1}\{Y_{it} = j\}}$$

where the product runs over the relevant cell: $(D_i = 0, t = 1)$ for $(\boldsymbol{\beta}_{01}, \boldsymbol{\xi}_{01})$, and $(D_i = 1, t = 0)$ for $(\boldsymbol{\beta}_{10}, \boldsymbol{\xi}_{10})$.

Finally, the counterfactual distribution is estimated as

$$\widehat{\Pr}(Y_{11} \leq j \mid \boldsymbol{X} = \boldsymbol{x}) = \widetilde{F}\left(\frac{\widehat{\kappa}_{j+1} - \widehat{\mu}_{11}(\boldsymbol{x})}{\widehat{\sigma}_{11}(\boldsymbol{x})} \right) \tag{C.4}$$

where

$$\widehat{\mu}_{11}(\boldsymbol{x}) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_{10} + \exp(\boldsymbol{x}^\top \widehat{\boldsymbol{\xi}}_{10}) \cdot \boldsymbol{x}^\top (\widehat{\boldsymbol{\beta}}_{01} - \widehat{\boldsymbol{\beta}}_{00}), \quad \text{and} \quad \widehat{\sigma}_{11}(\boldsymbol{x}) = \exp(\boldsymbol{x}^\top (\widehat{\boldsymbol{\xi}}_{10} + \widehat{\boldsymbol{\xi}}_{01})).$$

C.3 Asymptotic results

The main result of this section is the $\sqrt{n}$-consistency and asymptotic normality of the semiparametric treatment effect estimator $\widehat{\zeta}_j$.

Proposition A.5 (Consistency and asymptotic normality of the semiparametric estimator with covariates)**.** *Under the identification conditions in Corollary A.1 and the regularity conditions in Assumptions A.9 and A.10 below, the estimator $\widehat{\zeta}_j$ is consistent for the treatment effect ζ_j: $\widehat{\zeta}_j \xrightarrow{p} \zeta_j$. Furthermore, the estimator $\widehat{\zeta}_j$ is asymptotically normal:*

$$\sqrt{n}(\widehat{\zeta}_j - \zeta_j) \xrightarrow{d} \mathcal{N}(0, V_{jj}).$$

The proof is in Section E.9. It proceeds in three steps, each handled by a separate result below. First, I apply the results in Liu and Yu (2024) to establish the $\sqrt{n}$-consistency and asymptotic normality of the first-stage estimators $(\widehat{\boldsymbol{\beta}}_{00}, \widehat{\kappa}_2)$, along with the convergence rate of $\widehat{F}$ (Proposition A.6). Second, using a linearization of the NPMLE (Lemma A.1) together with profile likelihood theory, I derive an asymptotic linear expansion for the second-stage parameters $(\widehat{\boldsymbol{\beta}}_{01}, \widehat{\boldsymbol{\xi}}_{01})$ and $(\widehat{\boldsymbol{\beta}}_{10}, \widehat{\boldsymbol{\xi}}_{10})$ (Proposition A.7). Third, since the treatment effect estimator $\widehat{\zeta}_j$ depends on $\widehat{F}$ through a nonlinear functional (Equation (C.4)), a Taylor expansion in $\boldsymbol{\theta}$ alone does not suffice. Lemma A.2 shows that the predicted probabilities admit an asymptotic linear representation by decomposing the estimation error into parametric, nonparametric, and cross terms.

C.3.1 First-stage asymptotics

Assumption A.9 (Conditions for first-stage estimators)**.** (F1) The parameter space for $(\boldsymbol{\beta}_{00}, \boldsymbol{\kappa})$ is compact, the true parameters lie in its interior, and the cutoffs are strictly ordered and separated by a positive margin.

(F2) F_0 is twice continuously differentiable on the interior of the support, with a strictly positive continuous density f_0 bounded away from zero.

(F3) The linear index $\boldsymbol{X}^\top\boldsymbol{\beta}$ admits a continuous density $g_0(u;\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$. For $\boldsymbol{\beta} = \boldsymbol{\beta}_{00}$, the random variable $[\mathbf{1}\{Y_{i0} = 0\} - F_0(-\boldsymbol{X}_i^\top\boldsymbol{\beta}_{00})]\, g_0(\boldsymbol{X}_i^\top\boldsymbol{\beta}_{00} - \kappa_2; \boldsymbol{\beta}_{00})/g_0(\boldsymbol{X}_i^\top\boldsymbol{\beta}_{00}; \boldsymbol{\beta}_{00})$ has a finite second moment.

(F4) The density $g_0(u;\boldsymbol{\beta})$ and the conditional expectations $\mathbb{E}[\boldsymbol{X}_{-1} \mid \boldsymbol{X}^\top\boldsymbol{\beta} = u]$ and $\mathbb{E}[\boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \mid \boldsymbol{X}^\top\boldsymbol{\beta} = u]$ are twice continuously differentiable with respect to u. The maps $\boldsymbol{\beta} \mapsto g_0(u;\boldsymbol{\beta})$, $\boldsymbol{\beta} \mapsto \mathbb{E}[\boldsymbol{X}_{-1} \mid \boldsymbol{X}^\top\boldsymbol{\beta} = u]$, and $\boldsymbol{\beta} \mapsto \mathbb{E}[\boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \mid \boldsymbol{X}^\top\boldsymbol{\beta} = u]$ are continuous for all u in the support and all $\boldsymbol{\beta}$.

(F5) The Hessian matrix $\boldsymbol{H}_{\beta_0}$ is of full rank, and the scalar $V_{\kappa_0} = \int f_0(u+\kappa_0)\, g_0(u)\, du$ is nonzero.

Under these conditions, the results of Liu and Yu (2024) apply to the first-stage estimators. Their Theorem 3.3 covers the joint NPMLE estimator for the ordered response with $J = 3$ categories, where F is estimated from the interval-censored data as described in Section C.2. The following proposition establishes the asymptotic properties of the base distribution estimator $\widehat{F}_n$.

Proposition A.6 (Liu and Yu, 2024, Theorem 3.3 and Lemmas S12–S13)**.** *Under Assumption A.9, the parametric components $(\widehat{\boldsymbol{\beta}}_{00}, \widehat{\kappa}_2)$ are $\sqrt{n}$-consistent and asymptotically normal. The NPMLE $\widehat{F}_n$ satisfies:*

$$\sup_{\beta\in\mathcal{B}} \left(\int |\widehat{F}_n(u;\beta) - F_0(u)|^2 dG_0(u) \right)^{1/2} = O_p(n^{-1/3} \log^2 n),$$

and

$$\sup_{\beta\in\mathcal{B}} \sup_{u\in[C_L,C_U]} |\widehat{F}_n(u;\beta) - F_0(u)| = o_p(1).$$

C.3.2 Second-stage asymptotics

Let $s_j(\boldsymbol{x};\boldsymbol{\theta},\kappa) = (\kappa_{j+1} - \mu_{11}(\boldsymbol{x};\boldsymbol{\theta}))/\sigma_{11}(\boldsymbol{x};\boldsymbol{\theta})$ denote the standardized cutoff index.

Assumption A.10 (Conditions for the second-stage estimators)**.** (S1) The class $\{\phi_\theta(\boldsymbol{\theta}_0, F, \kappa_0) - \phi_\theta(\boldsymbol{\theta}_0, F_0, \kappa_0) \mid \|F - F_0\|_{L_2} \le \epsilon\}$ is P-Donsker, where $\phi_\theta(\boldsymbol{\theta}, F, \kappa)$ is defined as

$$\phi_\theta(\boldsymbol{\theta}, F, \kappa) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; F, \kappa), \quad \ell(\boldsymbol{\theta}; F, \kappa) = \sum_{j=0}^{J-1} \mathbf{1}\{Y_{it} = j\} \log \left(\widetilde{F}(s_{j+1}(\boldsymbol{X}_i; \boldsymbol{\theta}, \kappa)) - \widetilde{F}(s_j(\boldsymbol{X}_i; \boldsymbol{\theta}, \kappa)) \right).$$

(S2) The score and the second derivative of the second-stage log-likelihood are Lipschitz continuous with respect to F:

$$\|\phi_\theta(\boldsymbol{\theta}_0, F, \kappa_0) - \phi_\theta(\boldsymbol{\theta}_0, F_0, \kappa_0)\|_{L_2} \le L \|F - F_0\|_{L_2},$$

and

$$\|\partial_\theta \phi_\theta(\boldsymbol{\theta}_0, F, \kappa_0) - \partial_\theta \phi_\theta(\boldsymbol{\theta}_0, F_0, \kappa_0)\|_{L_2} \le L \|F - F_0\|_{L_2},$$

where $\|\cdot\|_{L_2}$ denotes the $L_2(P)$ norm with respect to the distribution of $(\boldsymbol{X}, Y)$.

(S3) The information matrix $H_{\boldsymbol{\theta}_0} \equiv -\mathbb{E}[\partial^2 \ell(\boldsymbol{\theta}_0; F_0, \kappa_0)/\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top]$ is positive definite.

(S4) For each j, the index $s_j(\boldsymbol{x}; \boldsymbol{\theta})$ is continuous in $\boldsymbol{x}$ and Lipschitz continuous in $\boldsymbol{\theta}$ uniformly over the support of $\boldsymbol{X}$. The parameter space Θ is compact and $\boldsymbol{\theta}_0$ lies in its interior.

(S5) The functional $F \mapsto P\{\phi_\theta(\boldsymbol{\theta}_0, F, \kappa_0)\}$ is pathwise differentiable at F_0, with derivative of the form $\int \psi_F \, d(F - F_0)$. The remainder is $O(\|F - F_0\|_{L_2}^2)$, which holds under condition F2 of Geskus and Groeneboom (1999).

Conditions (S1)–(S5) are standard in the semiparametric inference literature. Condition (S5) is verified for the interval-censoring NPMLE under the smoothness requirement in Assumption A.9(F2).

A key step in the proof of the second-stage expansion is the linearization of Stieltjes integrals against $\widehat{F} - F_0$. The following lemma, which relies on the canonical gradient for the NPMLE under interval censoring (Geskus and Groeneboom, 1999, Theorem 3.2), provides this representation.

Lemma A.1 (Linearization)**.** *The first-stage NPMLE $\widehat{F}$ satisfies*

$$\sqrt{n} \int c_j(u) \, d(\widehat{F}(u) - F_0(u)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_F(\boldsymbol{Z}_i; c_j) + o_p(1)$$

where $\phi_F(\cdot; c_j)$ is the influence function for $\widehat{F}$.

Proof is in Appendix E.7.

Using the linearization together with a mean-value expansion of the profile score around $(\boldsymbol{\theta}_0, F_0)$, we obtain the asymptotic linear representation of $\widehat{\boldsymbol{\theta}}$. The proof decomposes the score perturbation into an empirical process term (controlled via the Donsker condition (S1)) and a population-level term (controlled via the pathwise differentiability condition (S5) and the linearization in Lemma A.1). The CDF regularization lemma (Lemma A.10) ensures that replacing $\widehat{F}$ with its regularized version $\widetilde{F}$ introduces only an $o_p(n^{-1/2})$ error.

Proposition A.7 (Asymptotic linear expansion of second stage parameters)**.** *Assume that the assumptions required for identification hold. In addition, assume that the conditions in Assumptions A.9 and A.10 hold. Then,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = H_{\boldsymbol{\theta}_0}^{-1} \{\mathbb{G}_n \phi_\theta(\boldsymbol{\theta}_0, F_0, \kappa_0) + A_\kappa \sqrt{n}(\widehat{\kappa} - \kappa_0) + \mathbb{G}_n \psi_F\} + o_p(1)$$

Proof is in Appendix E.8.

C.3.3 Linear expansion of predicted probabilities

The remaining ingredient is to show that plugging both $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$ into the predicted counterfactual probabilities preserves the $\sqrt{n}$-rate. For each j, define the weight function

$$c_j(u) \equiv \Pr(s_j(\boldsymbol{X};\boldsymbol{\theta}_0) \le u \le s_{j+1}(\boldsymbol{X};\boldsymbol{\theta}_0) \mid D=1), \tag{C.5}$$

which is of bounded variation and has an essentially bounded Radon–Nikodym derivative with respect to f_0. This function arises as the density of the interval-censoring weights evaluated at the true parameter, and it governs how the nonparametric component $\widehat{F}$ enters the predicted probabilities.

Lemma A.2 (Linear expansion of predicted probabilities under the estimated base distribution)**.** *Assume that the conditions in Assumptions A.9 and A.10 hold. Let* $h_j(\boldsymbol{X};\boldsymbol{\theta},F) = F(s_{j+1}(\boldsymbol{X};\boldsymbol{\theta})) - F(s_j(\boldsymbol{X};\boldsymbol{\theta}))$*, and* $\widehat{G}_j(\boldsymbol{\theta},F) = \sum_{i=1}^n D_i\, h_j(\boldsymbol{X}_i;\boldsymbol{\theta},F)/n_1$*. Then,*

$$\widehat{G}_j(\widehat{\boldsymbol{\theta}},\widehat{F}) = \widehat{G}_j(\boldsymbol{\theta}_0,F_0) + \dot{G}_{j,\theta}(\boldsymbol{\theta}_0,F_0)^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \int c_j(u)\, d(\widehat{F}-F_0)(u) + o_p(n^{-1/2})$$

where $\dot{G}_{j,\theta}(\boldsymbol{\theta}_0,F_0) = \partial\widehat{G}_j(\boldsymbol{\theta},F)/\partial\boldsymbol{\theta}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0,F=F_0}$.

Proof is in Appendix E.10. The proof decomposes $\widehat{G}_j(\widehat{\boldsymbol{\theta}},\widehat{F}) - \widehat{G}_j(\boldsymbol{\theta}_0,F_0)$ into three terms: I_1 (Taylor expansion in $\boldsymbol{\theta}$ at fixed F_0), I_2 (Stieltjes integral in $\widehat{F}-F_0$ at fixed $\boldsymbol{\theta}_0$), and a cross term I_3. The first two terms yield the stated expansion, and the cross term is shown to be $o_p(n^{-1/2})$ by Lemma A.11.

Because $\widehat{G}_j$ averages over the n_1 treated observations, the natural empirical process in the proof is $\mathbb{G}_{n_1} = \sqrt{n_1}(\widehat{P}_{n_1} - P)$. Since $n_1/n \to p_1 \in (0,1)$ by Assumption A.5, all $o_p(n_1^{-1/2})$ remainders are also $o_p(n^{-1/2})$. The factor p_1 enters the asymptotic variance V_{jj} through the sampling variability of $\widehat{G}_j$.

C.4 Finite-sample performance

I evaluate the finite-sample performance of the semiparametric estimator through a Monte Carlo study. The data-generating process follows the location-scale model in (C.1) with $J=3$ categories, covariates $\boldsymbol{X} = (1, X_1, X_2)^\top$ where $X_1, X_2 \sim N(0,1)$, and the parameter values described in Section C.2 ($\delta = 0.3$, heteroskedastic scale model with $\boldsymbol{\xi}_{01}$ and $\boldsymbol{\xi}_{10}$ nonzero). I consider three error distributions: (i) $U_i \sim N(0,1)$, under which the ordered probit model is correctly specified; (ii) $U_i = T_i/\sqrt{5/3}$ where $T_i \sim t(5)$, a symmetric heavy-tailed distribution; and (iii) $U_i = (W_i - \mu_W)/\sigma_W$ where $W_i \sim 0.6 \cdot N(-1, 0.5^2) + 0.4 \cdot N(1.5, 1^2)$, a skewed mixture of normals ($\mu_W = 0$, $\sigma_W = \sqrt{2.05}$, skewness ≈ 0.72). The $t(5)$ distribution tests whether symmetric heavy tails cause probit misspecification, while the mixture error tests whether skewness does. Because the probit link Φ is symmetric, the ordered probit can absorb symmetric departures (such as heavy tails) through its location-scale parameters, but cannot capture skewness. The total sample size is $n \in \{500, 1000, 2000\}$, split equally between treated and control units.

Three estimators are compared: the parametric estimator (using standard normal distribution as the base CDF), the semiparametric NPMLE estimator with I-spline smoothing, and the semi-

parametric NPMLE estimator with the ridge regularization in (C.3). I use the bootstrap with $B = 200$ draws for computing standard errors. I run $R = 500$ Monte Carlo replications for each setting. Table 1 reports absolute bias, root mean squared error (RMSE), and coverage of 95% bootstrap confidence intervals, each averaged over the $J - 1 = 2$ treatment effect components ζ_1 and ζ_2.

Table 1: Finite-sample performance of the parametric (probit) and semiparametric (NPMLE) estimators with I-spline smoothing and ridge regularization, under correctly specified (normal) and misspecified ($t(5)$ and skewed mixture) error distributions. Absolute bias, RMSE, and coverage of 95% bootstrap percentile confidence intervals are averaged over $J - 1 = 2$ treatment effect components.

DGP	Method	n	Abs. Bias	RMSE	Coverage	SE/SD
Normal	Probit	500	0.003	0.051	94.4%	0.99
	Probit	1000	0.000	0.036	94.2%	0.98
	Probit	2000	0.001	0.025	94.2%	0.99
	NPMLE (spline)	500	0.007	0.057	97.6%	1.07
	NPMLE (spline)	1000	0.003	0.040	96.8%	1.05
	NPMLE (spline)	2000	0.003	0.028	96.1%	1.02
	NPMLE (ridge)	500	0.014	0.060	99.7%	1.04
	NPMLE (ridge)	1000	0.014	0.049	99.8%	1.06
	NPMLE (ridge)	2000	0.013	0.041	99.9%	1.07
$t(5)$	Probit	500	0.003	0.051	93.9%	0.96
	Probit	1000	0.002	0.037	93.7%	0.94
	Probit	2000	0.001	0.025	94.7%	1.01
	NPMLE (spline)	500	0.003	0.059	97.1%	1.03
	NPMLE (spline)	1000	0.004	0.041	96.6%	1.03
	NPMLE (spline)	2000	0.004	0.027	96.8%	1.07
	NPMLE (ridge)	500	0.008	0.056	99.9%	1.13
	NPMLE (ridge)	1000	0.015	0.051	99.8%	1.07
	NPMLE (ridge)	2000	0.013	0.042	99.5%	1.07
Mixture	Probit	500	0.005	0.045	92.6%	0.96
	Probit	1000	0.005	0.031	93.3%	0.99
	Probit	2000	0.004	0.022	92.6%	0.98
	NPMLE (spline)	500	0.007	0.059	98.4%	1.01
	NPMLE (spline)	1000	0.004	0.037	98.1%	1.08
	NPMLE (spline)	2000	0.003	0.026	97.5%	1.04
	NPMLE (ridge)	500	0.013	0.065	99.8%	1.06
	NPMLE (ridge)	1000	0.008	0.055	99.9%	1.02
	NPMLE (ridge)	2000	0.006	0.040	100.0%	1.16

Under correctly specified normal errors, all three estimators have low bias and RMSE that decrease at the expected $\sqrt{n}$ rate, and the bootstrap confidence intervals achieve approximately nominal coverage. The ordered probit estimator is slightly more efficient, as expected under correct specification. Under the $t(5)$ errors, the ordered probit performs nearly as well as under normal errors: the symmetric heavy tails are absorbed by the location-scale parameterization, producing

negligible bias. Under the skewed mixture errors, the ordered probit estimator exhibits persistent bias regardless of sample size, because Φ is symmetric and cannot capture the asymmetry of the true error distribution. The semiparametric NPMLE estimators maintain low bias under all three DGPs, confirming the robustness predicted by the theory. The I-spline smoothing variant achieves somewhat lower RMSE than the ridge regularization, because the smooth CDF improves numerical performance of the profile likelihood.

D Details on the Staggered Adoption Design

D.1 Setup

As before, we observe the outcome Y_{it} for n units. However, now we observe the outcome for $T+1$ time periods, $t = 0, 1, \ldots, T$. The treatment is administered at different time periods for different units. Let $D_{it} \in \{0, 1\}$ be the treatment indicator for unit i at time t, where $D_{it} = 1$ if unit i is treated at time t and $D_{it} = 0$ otherwise. I define $G_i \in \mathcal{G} \subseteq \{1, \ldots, T, \infty\}$ to denote the time period when unit i first receives the treatment, such that $G_i = \min\{t \mid D_{it} = 1\}$. I use $G_i = \infty$ to denote the units that never receive the treatment during the study period. The staggered adoption setup assumes that once a unit receives the treatment, it remains treated in all subsequent periods such that $D_{it} = \mathbf{1}\{t \geq G_i\}$.

D.2 Estimands

We are interested in estimating the following group-time specific effects at time $t \geq g$:

$$\begin{aligned}\zeta_j(g,t) &= \Pr(Y_{it}(1) = j \mid G_i = g) - \Pr(Y_{it}(0) = j \mid G_i = g) \\ \tau(g,t) &= \Pr(Y_{it}(1) > Y_{it}(0) \mid G_i = g) - \Pr(Y_{it}(1) < Y_{it}(0) \mid G_i = g).\end{aligned}$$

and overall effects that average over groups and time periods:

$$\overline{\zeta}_j = \sum_{g \in \mathcal{G}} \sum_{t \geq g} w_{gt} \zeta_j(g,t), \quad \text{and} \quad \overline{\tau} = \sum_{g \in \mathcal{G}} \sum_{t \geq g} w_{gt} \tau(g,t)$$

where $w_{gt} > 0$ are some user-specified weights that sum to one $\sum_g \sum_{t \geq g} w_{gt} = 1$. The choice of weights depends on the empirical application. Callaway and Sant'Anna (2021) provide a discussion on various types of aggregated effects and corresponding weights.

D.3 Identification

Let $F_{gt}(y) = \Pr(Y_{it}^*(0) \leq y \mid G_i = g)$ denote the CDF of the latent variable for unit i at time t for group $G_i = g$. Let $q_g^{s \to t}(v) = F_{gs}(F_{gt}^{-1}(v))$ denote the quantile-quantile transform from time s to t for the group that first receives the treatment at time g.

Assumption A.11 (Staggered distributional parallel trends)**.** Assume that the never-treated group exists in the data such that $\sum_{i=1}^n \mathbf{1}\{G_i > T\} > 0$. Then, for a fixed pair of indices (s, t) such that $s < g \leq t$

$$q_g^{s \to t}(v) = q_\infty^{s \to t}(v), \quad \forall v \in (0, 1).$$

Assumption A.11 requires that the quantile-quantile transform between time s and t is identical between group g and the never-treated group. This is a natural extension of the distributional parallel trends assumption (Assumption 3) to the staggered adoption setup. This is analogous to the group-specific PT assumption discussed in Callaway and Sant'Anna (2021). Note that the assumption that the never-treated group exists in the data is imposed for simplicity. When the never-treated group does not exist in the data, we can choose a set of units that are yet to be treated at time g as a comparison group.

Under Assumption A.11 and the location-scale model, we can identify the group-time specific treatment effects $\zeta_j(g,t)$ and $\tau(g,t)$ using a similar logic as before. Specifically, for $t \geq g$, we have

$$\Pr(Y_{it}(0) \leq j \mid G_i = g) = F_U\left(\frac{\kappa_j - \mu_{gt}}{\sigma_{gt}}\right)$$

where for $s < g \leq t$,

$$\mu_{gt} = \mu_{gs} + \sigma_{gs}\left(\frac{\mu_{\infty,t} - \mu_{\infty,s}}{\sigma_{\infty,s}}\right), \quad \text{and} \quad \sigma_{gt} = \sigma_{gs}\left(\frac{\sigma_{\infty,t}}{\sigma_{\infty,s}}\right).$$

D.4 Estimation

For the staggered adoption design, I assume the following location-scale family:

$$Y^*_{it}(0) \mid G_i = g \sim \mu_{gt} + \sigma_{gt} U_{it}$$

The identification of $\zeta_j(g,t)$ follows by applying the result in Proposition 1 by replacing $D_i = 1$ with $G_i = g$ and $D_i = 0$ with $G_i = \infty$. The bounds on $\tau(g,t)$ can be obtained similarly by applying the formula in Proposition A.1.

The estimation step requires some modifications. To simplify the notation, I continue to assume that there are never-treated units in the sample so that $\sum_{i=1}^n \mathbf{1}\{G_i = \infty\} > 0$, and use the first period of the never-treated group for normalization. Again, following Athey and Imbens (2006), I use Y_{gt} to denote the potential outcome for the group $G_i = g$ at time t such that $Y_{gt} \sim Y_{it}(0) \mid G_i = g$.

- Step 1: Estimate the parameters for $Y_{\infty,0}$ by fixing $\sigma_{\infty,0} = 1$ and $\kappa_1 = 0$.

$$(\widehat{\mu}_{\infty,0}, \widehat{\boldsymbol{\kappa}}) = \arg\max \sum_{i=1}^n \sum_{j \in \mathcal{J}} \mathbf{1}\{G_i = \infty, Y_{i0} = j\} \log\left\{F(\kappa_{j+1} - \mu_{\infty,0}) - F(\kappa_j - \mu_{\infty,0})\right\}.$$

- Step 2: Estimate the parameters for Y_{gt} for each $t < g$, while fixing the cutoffs $\widehat{\boldsymbol{\kappa}}$ obtained from Step 1.

$$(\widehat{\mu}_{gt}, \widehat{\sigma}_{gt}) = \arg\max \sum_{i:G_i=g} \sum_{j \in \mathcal{J}} \mathbf{1}\{Y_{it} = j\} \log\left\{F\left(\frac{\widehat{\kappa}_{j+1} - \mu}{\sigma}\right) - F\left(\frac{\widehat{\kappa}_j - \mu}{\sigma}\right)\right\}.$$

- Step 3: Estimate the counterfactual distribution of Y_{gt} for each g and t such that $g \leq t$

(post-treatment periods for $G_i = g$).

$$\widehat{\Pr}(Y_{gt} \leq j) = F\left(\frac{\widehat{\kappa}_{j+1} - \widehat{\mu}_{gt}}{\widehat{\sigma}_{gt}}\right)$$

where

$$\widehat{\mu}_{gt} = \widehat{\mu}_{gs} + \widehat{\sigma}_{gs} \cdot \frac{\widehat{\mu}_{\infty,t} - \widehat{\mu}_{\infty,s}}{\widehat{\sigma}_{\infty,s}}, \quad \text{and} \quad \widehat{\sigma}_{gt} = \widehat{\sigma}_{g,s} \cdot \widehat{\sigma}_{\infty,t}/\widehat{\sigma}_{\infty,s}.$$

for some $s < g$.

E Proofs of Main Results

E.1 Supporting Lemmas

Before proving propositions, we present useful lemmas.

Lemma A.3 (Identification of mean for Y_{00})**.** *Under Assumptions 1 and 2, with $\sigma_{00} = 1$ and $\kappa_1 = 0$, μ_{00} and all the other interior cutoffs $\{\kappa_j\}_{j=2}^{J-1}$ are identified as*

$$\mu_{00} = -F_U^{-1}(\Pr(Y_{00} = 0))$$

and

$$\kappa_j = \mu_{00} + F_U^{-1}\Big(\Pr(Y_{00} \leq j)\Big)$$

Proof. With $\sigma_{00} = 1$ and $\kappa_1 = 0$, we have $\Pr(Y_{00} \leq j) = F_U(\kappa_{j+1} - \mu_{00})$. Since F_U is invertible by Assumption 2, we obtain the desired result by inverting F_U. □

Lemma A.4 (Identification of parameters for Y_{01} and Y_{10})**.** *Under Assumptions 1 and 2, with fixed cutoffs $\boldsymbol{\kappa}$, the pair (μ_{01}, σ_{01}) is identified from the distribution of Y_{01}.*

Proof. Let $s_j = \Pr(Y_{01} \leq j)$ denote the cumulative probability for Y_{01} for each $j = 0, \ldots, J-1$. Then, we have

$$s_j = F_U\left(\frac{\kappa_{j+1} - \mu_{01}}{\sigma_{01}}\right)$$

If we have two categories j and j' such that $0 < s_j < 1$ and $0 < s_{j'} < 1$ and $s_j \neq s_{j'}$, we can solve the following two equations with two unknowns (μ_{01}, σ_{01}):

$$F_U^{-1}(s_j) = \frac{\kappa_{j+1} - \mu_{01}}{\sigma_{01}}$$
$$F_U^{-1}(s_{j'}) = \frac{\kappa_{j'+1} - \mu_{01}}{\sigma_{01}}$$

which gives the unique solution for (μ_{01}, σ_{01}).

Since F_U is strictly increasing by Assumption 2, the existence of such j and j' is guaranteed as long as $J \geq 3$.

The same argument applies to (μ_{10}, σ_{10}). This completes the proof. □

Lemma A.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 via MLE. Then, under regularity conditions similar to Assumption A.12, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is consistent to the true value $\boldsymbol{\theta}_0$ and asymptotically normal with variance-covariance matrix Ω,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \Omega) \tag{E.1}$$

Proof of Lemma A.5. The result is a direct application of the standard result of the maximum likelihood estimation. The proof is identical to that of Proposition 2 and therefore omitted. □

Lemma A.6 (Asymptotic Distribution of the Test Statistic)**.** *Let $r(v;\boldsymbol{\theta}) = \widetilde{q}_1(v;\boldsymbol{\theta}) - \widetilde{q}_0(v;\boldsymbol{\theta})$ and $\widehat{r}(v) \equiv r(v;\widehat{\boldsymbol{\theta}})$. Assume that $r(v;\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$ on $(0,1)\times\Theta$, and $\sup_{(v,\boldsymbol{\theta})\in(0,1)\times\Theta} \|\partial_{\boldsymbol{\theta}} r(v;\boldsymbol{\theta})\|_\infty < C < \infty$ for some neighborhood of $\boldsymbol{\theta}$ containing the true value. Then, we have that*

$$\sqrt{n}(\widehat{r}(v) - r(v)) \xrightarrow{d} \mathcal{N}\big(0, \mathrm{Var}(\widehat{r}(v))\big)$$

for each $v \in (0,1)$ with

$$\mathrm{Var}(\widehat{r}(v)) = \left(\frac{\partial}{\partial \boldsymbol{\theta}} r(v;\boldsymbol{\theta})\right)^\top \Omega \left(\frac{\partial}{\partial \boldsymbol{\theta}} r(v;\boldsymbol{\theta})\right)$$

where $\boldsymbol{\theta}$ is evaluated at the truth $\boldsymbol{\theta}_0$, Ω is the asymptotic variance covariance matrix of $\widehat{\boldsymbol{\theta}}$ given in Lemma A.5.

Proof of Lemma A.6. The result is a direct application of the Delta method. □

Lemma A.7 (Uniform convergence of Δ_n)**.** *Let $\Delta_n \equiv \sup_{v'} |\widehat{r}(v') - r(v')| + z_{1-\alpha} \sup_{v''} \widehat{\sigma}(v'')$ where $\widehat{\sigma}(v) = \sqrt{\widehat{V}(v)/n}$ with $\widehat{V}(v) = \widehat{\mathrm{Var}}(\widehat{r}(v))$. Assume that the condition in Lemma A.6 holds. Furthermore, assume that*

$$\sup_{v\in(0,1)} \left|\widehat{V}(v) - V(v)\right| = o_p(1).$$

Then, $\Delta_n = o_p(1)$.

Proof. Let $\mathcal{V} = (0,1)$. For each $v \in \mathcal{V}$, we apply the mean value theorem to obtain

$$\widehat{r}(v) - r(v) = \partial_\theta r(v;\bar{\theta})^\top (\widehat{\theta} - \theta_0)$$

where $\bar{\theta}$ lies between $\widehat{\theta}$ and θ_0. Then, we have

$$\begin{aligned}\sup_{v\in\mathcal{V}} |\widehat{r}(v) - r(v)| &\le \left(\sup_{(v,\theta)\in\mathcal{V}\times\Theta} \|\partial_\theta r(v;\theta)\|_\infty\right) \|\widehat{\theta} - \theta_0\| \\ &\le O(1)\cdot o_p(1) = o_p(1)\end{aligned}$$

where the first inequality is due to the condition in Lemma A.6, and the last line follows from the consistency of $\widehat{\theta}$ in Lemma A.5. This shows that the first term of Δ_n is $o_p(1)$.

Next, consider the second term of Δ_n.

$$\sup_{v\in\mathcal{V}} \widehat{\sigma}(v) = \frac{1}{\sqrt{n}} \sqrt{\sup_{v\in\mathcal{V}} \widehat{V}(v)}$$

$$\leq \frac{1}{\sqrt{n}}\left(\sup_{v\in\mathcal{V}}|\widehat{V}(v)-V(v)|+\sup_{v\in\mathcal{V}}V(v)\right)^{1/2}$$
$$\leq O(1/\sqrt{n})+o_p(1/\sqrt{n})=o_p(1)$$

where the second line follows from the triangle inequality. The uniform boundedness $\sup_v V(v)<\infty$ holds because $V(v)=\partial_\theta r(v;\theta_0)^\top\Omega\,\partial_\theta r(v;\theta_0)\leq C^2\|\Omega\|_{\mathrm{op}}$ under the conditions of Lemma A.6. This shows that the second term of Δ_n is $o_p(1)$, which completes the proof. □

Lemma A.8 (Validity of $(1-\alpha)$ level sets (Liu et al., 2009))**.** *Let $r(v)$ be a real-valued function defined on $v\in[0,1]$. Suppose $U_{1-\alpha}(v)$ is a point-wise upper $(1-\alpha)$ level confidence interval for the estimate $r(v)$ such that $\mathbb{P}(r(v)\leq U_{1-\alpha}(v))\geq 1-\alpha+o(1)$. Then,*

$$\mathbb{P}\left(\sup_{v\in(0,1)}r(v)\leq\sup_{v'\in(0,1)}U_{1-\alpha}(v')\right)\geq 1-\alpha+o(1).$$

Proof of Lemma A.8. Let $v^*\in\arg\max_v r(v)$. Then, we have that

$$\mathbb{P}(r(v^*)\leq U_{1-\alpha}(v^*))\leq\mathbb{P}\left(r(v^*)\leq\sup_{v'}U_{1-\alpha}(v')\right)$$

Then, the statement follows by realizing that the left hand side of the above inequality is at least $1-\alpha+o(1)$ by assumption. □

Lemma A.9 (Differentiability of min/max of finitely many linear maps)**.** *Under Assumption A.13, the function $g_U(p)=\min_{\ell\in\mathcal{L}}a_\ell^\top p$ is Hadamard differentiable at p with gradient $\nabla g_U(p)=a_{\ell_\star}$. Likewise, $g_L(p)=\max_{\ell\in\mathcal{L}'}b_\ell^\top p$ is differentiable at p with gradient $\nabla g_L(p)=b_{\ell^\dagger}$.*

Proof. Because the index sets are finite and each map is linear, the margin condition implies that in a neighborhood of p the same index remains active. Therefore, we have $g_U(p)=a_{\ell_\star}^\top p$ locally. The claim for g_L is analogous. □

Lemma A.10 (CDF regularization)**.** *Let $\varepsilon_n=n^{-1}$. Define the regularized CDF*

$$\widetilde{F}(u)=(1-2\varepsilon_n)\,\widehat{F}_n(u)+\varepsilon_n+\varepsilon_n\frac{u-\underline{u}}{\overline{u}-\underline{u}},$$

where $[\underline{u},\overline{u}]$ is the support of $\widehat{F}_n$. Then:

(i) *$\widetilde{F}$ is strictly increasing on $[\underline{u},\overline{u}]$*

(ii) *$\widetilde{F}(u)\in[\varepsilon_n,\,1]$ for all $u\in[\underline{u},\overline{u}]$*

(iii) *$\|\widetilde{F}-\widehat{F}_n\|_\infty\leq 2\varepsilon_n=O(n^{-1})=o(n^{-1/2})$.*

Proof. Write $\widetilde{F}(u)-\widehat{F}_n(u)=-2\varepsilon_n\widehat{F}_n(u)+\varepsilon_n+\varepsilon_n(u-\underline{u})/(\overline{u}-\underline{u})$. Since $\widehat{F}_n(u)\in[0,1]$ and $(u-\underline{u})/(\overline{u}-\underline{u})\in[0,1]$, the right-hand side lies in $[-\varepsilon_n,\,2\varepsilon_n]$. Hence $|\widetilde{F}(u)-\widehat{F}_n(u)|\leq 2\varepsilon_n$ for every u, giving (iii). Property (i) holds because $\widehat{F}_n$ is non-decreasing and the linear term $\varepsilon_n(u-\underline{u})/(\overline{u}-\underline{u})$ is strictly increasing. Property (ii) follows from $\widetilde{F}(\underline{u})=(1-2\varepsilon_n)\widehat{F}_n(\underline{u})+\varepsilon_n\geq\varepsilon_n$ and $\widetilde{F}(u)\leq(1-2\varepsilon_n)+2\varepsilon_n=1$. □

Lemma A.11 (Cross-term negligibility)**.** *Under Assumptions A.9 and A.10, with $\widehat{G}_j$ and c_j as defined in Lemma A.2, the cross term from the decomposition in its proof,*

$$I_3 \equiv \widehat{G}_j(\widehat{\boldsymbol{\theta}}, \widehat{F}) - \widehat{G}_j(\widehat{\boldsymbol{\theta}}, F_0) - \widehat{G}_j(\boldsymbol{\theta}_0, \widehat{F}) + \widehat{G}_j(\boldsymbol{\theta}_0, F_0),$$

satisfies $I_3 = o_p(n^{-1/2})$.

Proof. We decompose I_3 by adding and subtracting the population analogue of $\widehat{c}_{j,\theta}(u)$:

$$\begin{aligned} I_3 &= \int \Big[\widehat{c}_{j,\widehat{\theta}}(u) - \widehat{c}_{j,\theta_0}(u)\Big] d(\widehat{F} - F_0)(u) \\ &= \int \Big[[\widehat{c}_{j,\widehat{\theta}}(u) - c_{j,\widehat{\theta}}(u)] - [\widehat{c}_{j,\theta_0}(u) - c_{j,\theta_0}(u)]\Big] d(\widehat{F} - F_0)(u) + \int \Big[c_{j,\widehat{\theta}}(u) - c_{j,\theta_0}(u)\Big] d(\widehat{F} - F_0)(u) \\ &\equiv I_3^{(a)} + I_3^{(b)}. \end{aligned}$$

The first piece $I_3^{(a)}$ involves the empirical process fluctuation of the weight functions, while $I_3^{(b)}$ involves only the population weights evaluated at $\widehat{\theta}$ versus θ_0. We bound each in turn.

For $I_3^{(a)}$, we apply the bound $|\int g\, d\nu| \le \|g\|_\infty \|\nu\|_{TV}$ together with $\|\widehat{F} - F_0\|_{TV} \le 2$:

$$\begin{aligned} |I_3^{(a)}| &\le \sup_{u \in \mathcal{U}} \big|[\widehat{c}_{j,\widehat{\theta}}(u) - c_{j,\widehat{\theta}}(u)] + [\widehat{c}_{j,\theta_0}(u) - c_{j,\theta_0}(u)]\big| \|\widehat{F} - F_0\|_{TV} \\ &\le 2 \sup_{u \in \mathcal{U}} \big|[\widehat{c}_{j,\widehat{\theta}}(u) - c_{j,\widehat{\theta}}(u)] + [\widehat{c}_{j,\theta_0}(u) - c_{j,\theta_0}(u)]\big|. \end{aligned}$$

To bound the supremum, write the centered weight function as an empirical process:

$$\begin{aligned} \sqrt{n_1}(\widehat{c}_{j,\theta}(u) - c_{j,\theta}(u)) &= \sqrt{n_1}(\mathbb{P}_{n_1} - P)\mathbf{1}\{s_j(\boldsymbol{X};\theta) \le u \le s_{j+1}(\boldsymbol{X};\theta)\} \\ &= \mathbb{G}_{n_1}\mathbf{1}\{s_j(\boldsymbol{X};\theta) \le u \le s_{j+1}(\boldsymbol{X};\theta)\} \equiv \mathbb{G}_{n_1} f_{j,u,\theta}. \end{aligned}$$

The difference of the centered weights at $\widehat{\theta}$ and θ_0 is then

$$\sqrt{n_1}\Big\{[\widehat{c}_{j,\widehat{\theta}}(u) - c_{j,\widehat{\theta}}(u)] + [\widehat{c}_{j,\theta_0}(u) - c_{j,\theta_0}(u)]\Big\} = \mathbb{G}_{n_1}(f_{j,u,\widehat{\theta}} - f_{j,u,\theta_0}).$$

Since the class $\{f_{j,u,\theta}\}$ is P-Donsker and $\widehat{\theta} \xrightarrow{p} \theta_0$, stochastic equicontinuity gives

$$\sqrt{n_1}|I_3^{(a)}| \le 2 \sup_{u \in \mathcal{U}} \big|\mathbb{G}_{n_1}\big(f_{j,u,\widehat{\theta}} - f_{j,u,\theta_0}\big)\big| = o_p(1)$$

and hence $I_3^{(a)} = o_p(n^{-1/2})$, using $n_1/n \to p_1 > 0$.

For $I_3^{(b)}$, the integrand involves only population weights. We can write

$$I_3^{(b)} = \int [c_{j,\widehat{\theta}}(u) - c_{j,\theta_0}(u)]\, d(\widehat{F} - F_0)(u).$$

By the Lipschitz continuity of $s_m(\boldsymbol{x};\theta)$ in θ (Assumption A.10(S4)) and the bounded density of F_0 (Assumption A.9(F2)), the population weight function satisfies

$$\|c_{j,\widehat{\theta}} - c_{j,\theta_0}\|_\infty \le C\|\widehat{\theta} - \theta_0\| = O_p(n^{-1/2}).$$

Since $\widehat{F} - F_0$ has bounded total variation ($\|\widehat{F} - F_0\|_{TV} \leq 2$), we have

$$|I_3^{(b)}| \leq \|c_{j,\widehat{\theta}} - c_{j,\theta_0}\|_\infty \cdot \|\widehat{F} - F_0\|_{TV} = O_p(n^{-1/2}) \cdot O(1) = O_p(n^{-1/2}).$$

To sharpen this to $o_p(n^{-1/2})$, we apply the linearization in Lemma A.1 with the shrinking weight $g_n = c_{j,\widehat{\theta}} - c_{j,\theta_0}$. The linearization gives $\int g_n \, d(\widehat{F} - F_0) = n^{-1} \sum_i \phi_F(\boldsymbol{Z}_i; g_n) + o_p(n^{-1/2})$. Since $\|g_n\|_\infty = O_p(n^{-1/2})$ and $\phi_F(\cdot; g_n)$ inherits the same rate, the sum is $O_p(n^{-1/2}) \cdot O_p(n^{-1/2}) = o_p(n^{-1/2})$. Combining the bounds on $I_3^{(a)}$ and $I_3^{(b)}$ completes the proof. □

E.2 Proof of Proposition 1

Proof. 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

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

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

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

By Assumption 3, we have

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

where

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

Combining the results in Lemmas A.3 and A.4, all the parameters (μ_{11}, σ_{11}) are identified with the normalization restriction that $\sigma_{00} = 1$ and $\kappa_1 = 0$. □

E.3 Proof of Proposition 2

Assumption A.12 (Regularity conditions)**.** (R1) $\{(Y_{i0}, Y_{i1}, D_i)\}_{i=1}^n$ are i.i.d. across i. Within-unit dependence across t is unrestricted, and $n_1/n \to p_1 \in (0,1)$.

(R2) The ordered index model with known F_U holds with strictly ordered cutoffs that are fixed across cells.

(R3) The per-unit log-likelihood contributions are twice continuously differentiable, the true $\boldsymbol{\theta}_0$ lies in the interior of a compact parameter space, the Fisher information exists and is nonsingular, score moments are finite, and a ULLN holds for the stacked score.

(R4) $r_j(\boldsymbol{\theta}) = F_U\big((\kappa_{j+1} - \mu_{11})/\sigma_{11}\big) - F_U\big((\kappa_j - \mu_{11})/\sigma_{11}\big)$ is continuously differentiable near $\boldsymbol{\theta}_0$.

The proof for the consistency and asymptotic normality of $\widehat{\zeta}$ and $\widehat{\Delta}$ directly follows from the standard M-estimation theory and the delta method. The proof is presented below for completeness.

Proof. The stacked score from the three likelihood blocks is

$$g_i(\theta) = \begin{bmatrix} (1 - D_i)\, \partial_{\mu_{00},\kappa}\, \ell_{00}(Y_{i0}; \mu_{00}, \kappa) \\ (1 - D_i)\, \partial_{\mu_{01},\sigma_{01}}\, \ell_{01}(Y_{i1}; \mu_{01}, \sigma_{01}, \kappa) \\ D_i\, \partial_{\mu_{10},\sigma_{10}}\, \ell_{10}(Y_{i0}; \mu_{10}, \sigma_{10}, \kappa) \end{bmatrix}.$$

Under (R1)–(R3), standard M-estimation theory gives consistency of $\widehat{\theta}$ and the asymptotic linear representation

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) \;=\; -A^{-1}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\theta_0) + o_p(1),$$

where $A = \mathbb{E}[(\partial/\partial\theta) g_i(\theta)|_{\theta=\theta_0}]$. Define the counterfactual probability as a function of θ: $r_j(\theta) = F_U((\kappa_{j+1} - \mu_{11})/\sigma_{11}) - F_U((\kappa_j - \mu_{11})/\sigma_{11})$. By (R4) and the delta method,

$$\sqrt{n}\,\big(\widehat{r}_j - r_j(\theta_0)\big) = -\,\partial_\theta r_j(\theta_0)^\top A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\theta_0) + o_p(1).$$

The post-treatment treated probability $\pi_j = \Pr(Y_{i1} = j \mid D_i = 1)$ is estimated by its empirical analogue $\widehat{\pi}_j$, which satisfies

$$\sqrt{n}\big(\widehat{\pi}_j - \pi_j\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{p_1}\big(1\{Y_{i1} = j\} - \pi_j\big) + o_p(1).$$

Since $\widehat{\zeta}_j = \widehat{\pi}_j - \widehat{r}_j$ and $\zeta_j = \pi_j - r_j(\theta_0)$, combining the two expansions yields

$$\sqrt{n}(\widehat{\zeta}_j - \zeta_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{ij} + o_p(1), \quad \psi_{ij} = \frac{D_i}{p_1}\big(\mathbf{1}\{Y_{i1} = j\} - \pi_j\big) - \partial_\theta r_j(\theta_0)^\top A^{-1} g_i(\theta_0).$$

A multivariate CLT gives the first claim with $V_{\mathcal{J}} = \mathbb{E}[\psi_{i,\mathcal{J}}\psi_{i,\mathcal{J}}^\top]$. The plug-in estimator $\widehat{V}_{\mathcal{J}} = (1/n)\sum_i \widehat{\psi}_{i,\mathcal{J}}\widehat{\psi}_{i,\mathcal{J}}^\top$ is consistent by the continuous mapping theorem, and the result for $\widehat{\Delta} = M\widehat{\zeta}$ follows by Slutsky's theorem. □

To prove the joint asymptotic normality of the lower and upper bound estimators, I impose an additional assumption on the margin condition.

Assumption A.13 (Margin for the active bound)**.** Let $\{\delta_\ell(p) = a_\ell^\top p : \ell \in \mathcal{L}\}$ denote the finite family of linear functionals that appear in Proposition A.1. There exists a unique $\ell_\star \in \mathcal{L}$ attaining

the minimum at the truth and a strictly positive gap

$$\Delta_U \equiv \min_{\ell \neq \ell_\star} \{\delta_\ell(p) - \delta_{\ell_\star}(p)\} > 0.$$

Similarly, for the lower bound $\gamma_L = \max_{\ell \in \mathcal{L}'} \xi_\ell(p) = \max_{\ell \in \mathcal{L}'} b_\ell^\top p$, there is a unique $\ell^\dagger \in \mathcal{L}'$ attaining the maximum with gap

$$\Delta_L \equiv \min_{\ell \neq \ell^\dagger} \{\xi_{\ell^\dagger}(p) - \xi_\ell(p)\} > 0.$$

Proof. From the previous step of the proof of this proposition, we have already established that $\sqrt{n}(\widehat{p} - p)$ converges in distribution to a multivariate normal distribution. By Lemma A.9 and the delta method, g_U and g_L are differentiable at p with gradients $a_{\ell_\star}$ and $b_{\ell^\dagger}$. Therefore,

$$\sqrt{n}\{g_U(\hat{p}) - g_U(p)\} = a_{\ell_\star}^\top \sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, a_{\ell_\star}^\top \Sigma_p a_{\ell_\star}),$$

and likewise for g_L. Consistency of the plug-in variance formulas follows from continuous mapping and consistency of $(\hat{G}, \hat{\Omega})$. □

E.4 Proof of Proposition 3

Proof. Under the null $\mathbb{P} \in H_0$, we have either H_0^+ or H_0^- holds.

Suppose that H_0^+ holds without loss of generality. Then,

$$\begin{aligned}
\mathbb{P}(\mathcal{R}_+ \cap \mathcal{R}_-) &\leq \mathbb{P}(\mathcal{R}_+) \\
&= \mathbb{P}\left(\sup_v U_{1-\alpha}(v) < \delta \right) \\
&\leq \mathbb{P}\left(\sup_v U_{1-\alpha}(v) < \sup_{v'} r(v') \right) \\
&= 1 - \mathbb{P}\left(\sup_v U_{1-\alpha}(v) \geq \sup_{v'} r(v') \right) \\
&\leq 1 - (1 - \alpha + o(1)) = \alpha + o(1)
\end{aligned}$$

where the third line holds because under H_0^+ we have $\delta \leq \sup_v r(v)$, and the last inequality uses Lemma A.8.

The case where H_0^- holds is symmetric. □

E.5 Proof of Proposition A.2

Proof. Consider the upper event.

$$\begin{aligned}
\sup_v U_{1-\alpha}(v) &= \sup_v \left\{ \widehat{r}(v) + z_{1-\alpha} \sqrt{\widehat{\mathrm{Var}}(\widehat{r}(v))/n} \right\} \\
&\leq \sup_v r(v) + \sup_{v'} |\widehat{r}(v') - r(v')| + z_{1-\alpha} \sup_{v''} \sqrt{\widehat{\mathrm{Var}}(\widehat{r}(v''))/n} \\
&\leq \delta - \epsilon + \Delta_n
\end{aligned}$$

where $\Delta_n \equiv \sup_{v'} |\widehat{r}(v') - r(v')| + z_{1-\alpha} \sup_{v''} \sqrt{\widehat{\mathrm{Var}}(\widehat{r}(v''))/n}$.

The above implies that it is sufficient to have $\delta - \epsilon + \Delta_n < \delta$ for the rejection event to hold. Therefore, we have

$$\begin{aligned}\Pr(\mathcal{R}_+) &= \Pr\left(\sup_v U_{1-\alpha}(v) < \delta\right) \\ &\geq \Pr(\Delta_n < \epsilon) \;\to\; 1\end{aligned}$$

where the last line follows from Lemma A.7. This proves that $\Pr(\mathcal{R}_+) \to 1$.

The lower event $\mathcal{R}_-$ can be handled similarly, and therefore I omit the proof.

Combining the two events, we have

$$P(\mathcal{R}_+ \cap \mathcal{R}_-) \to 1$$

which completes the proof. □

E.6 Proof of Proposition A.3

Proof. Let $\widetilde{F}_{11}(r) = q_0^{-1} F_{10}(y)$ denote the counterfactual CDF of Y_{11}^* obtained under the distributional parallel-trends assumption.

By the definition of q_1, we have

$$F_{11}^*(y) = q_1^{-1}(F_{10}(y)).$$

where $F_{11}^*(y)$ is the true CDF of Y_{11}^*.

Now, consider the difference between the true effect $\zeta_j = F_{11}^*(\kappa_{j+1}) - F_{11}^*(\kappa_j)$ and $\widetilde{\zeta}_j = \widetilde{F}_{11}(\kappa_{j+1}) - \widetilde{F}_{11}(\kappa_j)$,

$$\begin{aligned}\zeta_j - \widetilde{\zeta}_j &= [q_1^{-1}(F_{10}(\kappa_{j+1})) - q_0^{-1}(F_{10}(\kappa_{j+1}))] - [q_1^{-1}(F_{10}(\kappa_j)) - q_0^{-1}(F_{10}(\kappa_j))] \\ &\leq 2 \sup_{v\in(0,1)} |q_1^{-1}(v) - q_0^{-1}(v)| \\ &\leq 2 \sup_{v\in(0,1)} \frac{|q_1(v) - q_0(v)|}{\inf_{v'\in(0,1)} q_0'(v')} \\ &= 2\delta/M\end{aligned}$$

where the third line follows from the mean value theorem and the last line follows from the definition of δ and M.

Finally, to obtain the final result, we note that

$$\begin{aligned}|\text{Bias}(\widehat{\zeta}_j)| &\leq |\widehat{\zeta}_j - \widetilde{\zeta}_j| + |\widetilde{\zeta}_j - \zeta_j| \\ &= |\widetilde{\zeta}_j - \zeta_j| + o_p(1) \\ &\leq 2\delta/M + o_p(1)\end{aligned}$$

where the first line follows from the triangle inequality, the second line follows from the consistency of $\widehat{\zeta}_j$, and the last line follows from the previous result.

The bound for Δ_j follows from the same argument. Since $\Delta_j = \Pr(Y_{i1}(0) \geq j \mid D_i = 1) -$

$\Pr(Y_{i1}(1) \geq j \mid D_i = 1)$ involves a single CDF evaluation at κ_j rather than the difference of two,

$$\begin{aligned}\Delta_j - \widetilde{\Delta}_j &= q_1^{-1}(F_{10}(\kappa_j)) - q_0^{-1}(F_{10}(\kappa_j)) \\ &\leq \sup_{v\in(0,1)} |q_1^{-1}(v) - q_0^{-1}(v)| \leq \delta/M,\end{aligned}$$

and the same consistency argument gives $|\text{Bias}(\widehat{\Delta}_j)| \leq \delta/M + o_p(1)$. □

E.7 Proof of Lemma A.1

Proof. The interval-censoring NPMLE $\widehat{F}$ admits an asymptotic linear representation via the canonical gradient for the NPMLE under interval censoring (Geskus and Groeneboom, 1999, Theorem 3.2). In particular, the canonical gradient for the interval-censoring NPMLE is derived in Geskus and Groeneboom (1999) (Theorem 3.2) and applied to the ordered response setting in Liu and Yu (2024) (eq. (3.18) and supplement (S.86)–(S.93)). For $J = 3$, this yields the asymptotic linear representation via Lemma S17 of Liu and Yu (2024). □

E.8 Proof of Proposition A.7

Proof. The solution $\widehat{\theta}$ satisfies the first-order condition:

$$\frac{1}{n}\sum_{i=1}^{n} \phi_\theta(\widehat{\theta}, \widetilde{F}, \widehat{\kappa}) = 0.$$

Applying a mean-value expansion around the truth $(\theta_0, F_0, \kappa_0)$,

$$\begin{aligned}0 &= \mathbb{P}_n\phi_\theta(\theta_0, \widetilde{F}, \kappa_0) + \mathbb{P}_n\partial_\theta\phi_\theta(\bar{\theta}, \widetilde{F}, \bar{\kappa})(\widehat{\theta} - \theta_0) + \mathbb{P}_n\partial_\kappa\phi_\theta(\bar{\theta}, \widetilde{F}, \bar{\kappa})(\widehat{\kappa} - \kappa_0) \\ &= \mathbb{P}_n\phi_\theta(\theta_0, F_0, \kappa_0) + P\partial_\theta\phi_\theta(\theta_0, F_0, \kappa_0)(\widehat{\theta} - \theta_0) + P\partial_\kappa\phi_\theta(\theta_0, F_0, \kappa_0)(\widehat{\kappa} - \kappa_0) \\ &\quad + \mathbb{P}_n\phi_\theta(\theta_0, \widetilde{F}, \kappa_0) - \mathbb{P}_n\phi_\theta(\theta_0, F_0, \kappa_0)\end{aligned}$$

where the second equality follows from the GC property and the Lipschitz continuity of the score and Hessian functions, which implies that

$$(\mathbb{P}_n - P)\partial_\theta\phi_\theta(\bar{\theta}, \widetilde{F}, \bar{\kappa}) = o_p(1)$$

and

$$|P\partial_\theta\phi_\theta(\bar{\theta}, \widetilde{F}, \bar{\kappa}) - P\partial_\theta\phi_\theta(\theta_0, F_0, \kappa_0)| = o_p(1).$$

The final term can be written as

$$\begin{aligned}\mathbb{P}_n\{\phi_\theta(\theta_0, \widetilde{F}, \kappa_0) - \phi_\theta(\theta_0, F_0, \kappa_0)\} &= (\mathbb{P}_n - P)\{\phi_\theta(\theta_0, \widetilde{F}, \kappa_0) - \phi_\theta(\theta_0, F_0, \kappa_0)\} + P\{\phi_\theta(\theta_0, \widetilde{F}, \kappa_0) - \phi_\theta(\theta_0, F_0, \kappa_0)\} \\ &\equiv I_n^a + I_n^b\end{aligned}$$

We control I_n^a using the Donsker property. Let $f_n = \phi_\theta(\theta_0, \widetilde{F}, \kappa_0) - \phi_\theta(\theta_0, F_0, \kappa_0)$. By Proposition A.6, $\|\widetilde{F} - F_0\|_{L_2} = o_p(1)$, so f_n lies in the Donsker class of Assumption A.10(S1) with

probability approaching one. Moreover, by the Lipschitz condition in (S2),

$$\|f_n\|_{L_2(P)} = \|\phi_\theta(\theta_0, \widetilde{F}, \kappa_0) - \phi_\theta(\theta_0, F_0, \kappa_0)\|_{L_2(P)} \leq L\|\widetilde{F} - F_0\|_{L_2} = o_p(1).$$

Since $f_n \to 0$ in $L_2(P)$ and f_n belongs to a P-Donsker class with probability tending to one, stochastic equicontinuity (Lemma 19.24 of van der Vaart, 2000) gives $\sqrt{n}\, I_n^a = \mathbb{G}_n f_n = o_p(1)$.

The final term I_n^b can be controlled using the pathwise differentiability condition:

$$\begin{aligned} I_n^b &= \int \psi_F \, d(\widetilde{F} - F_0) + O(\|\widetilde{F} - F_0\|^2) \\ &= \int \psi_F \, d(\widehat{F} - F_0) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \psi_F(\boldsymbol{Z}_i) + o_p(n^{-1/2}) \end{aligned}$$

where the first equality follows from Assumption A.10(S5), and the second equality follows from the convergence rate of $\widehat{F}$ in Proposition A.6 and Lemma A.10, and the final equality follows from the linearization of $\widehat{F}$ in Lemma A.1. Thus, it follows that

$$\sqrt{n} I_n^{(b)} = \mathbb{G}_n \psi_F + o_p(1).$$

Finally, by rearranging the terms, we have

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &= -\{P\partial_\theta \phi_\theta(\theta_0, F_0, \kappa_0)\}^{-1}\{\mathbb{G}_n \phi_\theta(\theta_0, F_0, \kappa_0) + P\partial_\kappa \phi_\theta(\theta_0, F_0, \kappa_0)(\widehat{\kappa} - \kappa_0) + \mathbb{G}_n \psi_F\} + o_p(1) \\ &= H_{\theta_0}^{-1}\{\mathbb{G}_n \phi_\theta(\theta_0, F_0, \kappa_0) + A_\kappa \sqrt{n}(\widehat{\kappa} - \kappa_0) + \mathbb{G}_n \psi_F\} + o_p(1) \end{aligned}$$

which concludes the proof. □

E.9 Proof of Proposition A.5

Proof. Let $s_j(\boldsymbol{x}; \theta) = (\kappa_j - \mu_{11}(\boldsymbol{x}; \theta))/\sigma_{11}(\boldsymbol{x}; \theta)$ and $h_j(\boldsymbol{x}; \theta, F) = F(s_{j+1}(\boldsymbol{x}; \theta)) - F(s_j(\boldsymbol{x}; \theta))$, where $\theta = (\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\kappa})$. Then, the counterfactual probability can be written as

$$\Pr(Y_{i1}(0) = j \mid D_i = 1) \equiv G_j(\theta, F) = \mathbb{E}[h_j(\boldsymbol{X}_i; \theta, F) \mid D_i = 1].$$

and its estimator is given by

$$\widehat{G}_j = \frac{1}{n_1}\sum_{i=1}^{n} h_j(\boldsymbol{X}_i; \widehat{\theta}, \widehat{F}).$$

For a given $\boldsymbol{x}$,

$$\begin{aligned} h_j(\boldsymbol{x}; \widehat{\theta}, \widehat{F}) - h_j(\boldsymbol{x}; \theta_0, F_0) &= \left[\widehat{F}(s_{j+1}(\boldsymbol{x}; \widehat{\theta})) - F_0(s_{j+1}(\boldsymbol{x}; \theta_0))\right] - \left[\widehat{F}(s_j(\boldsymbol{x}; \widehat{\theta})) - F_0(s_j(\boldsymbol{x}; \theta_0))\right] \\ &= \left[\widehat{F}(s_{j+1}(\boldsymbol{x}; \widehat{\theta})) - \widehat{F}(s_{j+1}(\boldsymbol{x}; \theta_0))\right] - \left[\widehat{F}(s_j(\boldsymbol{x}; \widehat{\theta})) - \widehat{F}(s_j(\boldsymbol{x}; \theta_0))\right] \\ &\quad + \left[\widehat{F}(s_{j+1}(\boldsymbol{x}; \theta_0)) - F_0(s_{j+1}(\boldsymbol{x}; \theta_0))\right] - \left[\widehat{F}(s_j(\boldsymbol{x}; \theta_0)) - F_0(s_j(\boldsymbol{x}; \theta_0))\right] \\ &\equiv [I_1^{(a)}] - [I_1^{(b)}] + [I_2^{(a)}] - [I_2^{(b)}] \end{aligned}$$

We bound each group of terms in turn. The terms $I_2^{(a)}$ and $I_2^{(b)}$ measure the approximation error of $\widehat{F}$ at the true index values, and are controlled by the uniform convergence of the NPMLE:

$$|I_2^{(a)} - I_2^{(b)}| \leq |I_2^{(a)}| + |I_2^{(b)}| \leq 2 \sup_{u \in \mathcal{U}} |\widehat{F}(u) - F_0(u)| = o_p(1)$$

where the last equality follows from Proposition A.6 on the uniform convergence of $\widehat{F}$ to F_0.

Next, consider $I_1^{(a)}$, which captures the effect of estimating θ. Adding and subtracting F_0 at the estimated index, we have

$$\begin{aligned}
I_1^{(a)} &= F_0(s_{j+1}(\boldsymbol{x}; \widehat{\theta})) - F_0(s_{j+1}(\boldsymbol{x}; \theta_0)) \\
&\quad + [\widehat{F}(s_{j+1}(\boldsymbol{x}; \widehat{\theta})) - F_0(s_{j+1}(\boldsymbol{x}; \widehat{\theta}))] - [\widehat{F}(s_{j+1}(\boldsymbol{x}; \theta_0)) - F_0(s_{j+1}(\boldsymbol{x}; \theta_0))] \\
&\leq f(s_{j+1}(\boldsymbol{x}; \bar{\theta}))[s_{j+1}(\boldsymbol{x}; \widehat{\theta}) - s_{j+1}(\boldsymbol{x}; \theta_0)] + 2 \sup_{u \in \mathcal{U}} |\widehat{F}(u) - F_0(u)| + o_p(1) \\
&\leq L \|\widehat{\theta} - \theta_0\|_2 + o_p(1) = o_p(1)
\end{aligned}$$

where the second inequality follows from the bounded density of F_0, and the last line follows from the Lipschitz continuity of $s_j(\boldsymbol{x}; \theta)$ (Assumption A.10), the application of the continuous mapping theorem on the result on Proposition A.7, and the uniform convergence of $\widehat{F}$ to F_0. The term $I_1^{(b)}$ is handled identically. Collecting the four bounds yields

$$\sup_{\boldsymbol{x} \in \mathcal{X}} |h_j(\boldsymbol{x}; \widehat{\theta}, \widehat{F}) - h_j(\boldsymbol{x}; \theta_0, F_0)| \leq C \|\widehat{\theta} - \theta_0\|_2 + 4 \sup_{u \in \mathcal{U}} |\widehat{F}(u) - F_0(u)| = o_p(1).$$

Combining this with the law of large numbers proves the consistency.

The asymptotic normality follows from Lemma A.2 and Proposition 2 on the estimated proportion on the observed treated outcomes. The application of the multivariate CLT concludes the proof. □

E.10 Proof of Lemma A.2

Proof. Adding and subtracting the terms, we have

$$\begin{aligned}
\widehat{G}_j(\widehat{\theta}, \widehat{F}) - \widehat{G}_j(\theta_0, F_0) &= \left[\widehat{G}_j(\widehat{\theta}, F_0) - \widehat{G}_j(\theta_0, F_0)\right] \\
&\quad + \left[\widehat{G}_j(\theta_0, \widehat{F}) - \widehat{G}_j(\theta_0, F_0)\right] \\
&\quad + \left[\widehat{G}_j(\widehat{\theta}, \widehat{F}) - \widehat{G}_j(\widehat{\theta}, F_0) - \widehat{G}_j(\theta_0, \widehat{F}) + \widehat{G}_j(\theta_0, F_0)\right] \\
&\equiv I_1 + I_2 + I_3
\end{aligned}$$

The first term I_1 isolates the effect of estimating θ while holding F fixed. By a Taylor expansion,

$$I_1 = \dot{G}_{j,\theta}(\theta_0, F_0)^\top (\widehat{\theta} - \theta_0) + o_p(\|\widehat{\theta} - \theta_0\|^2).$$

where

$$\dot{G}_{j,\theta}(\theta_0, F_0) = \left. \frac{\partial \widehat{G}_j(\theta, F)}{\partial \theta} \right|_{\theta = \theta_0, F = F_0}.$$

The remainder term is $o_p(n^{-1/2})$ by the $\sqrt{n}$-consistency of $\widehat{\theta}$ which follows by Delta method and

Proposition A.7. This shows that

$$I_1 = \dot{G}_{j,\theta}(\theta_0, F_0)^\top(\widehat{\theta} - \theta_0) + o_p(n^{-1/2}).$$

The second term I_2 captures the effect of replacing F_0 with $\widehat{F}$ at the true parameter. For notational simplicity, let $\widehat{c}_{j,\theta}(u)$ be defined as

$$\widehat{c}_{j,\theta}(u) = \frac{1}{n_1}\sum_{i=1}^{n} D_i \mathbf{1}\{s_j(\boldsymbol{X}_i;\theta) \le u \le s_{j+1}(\boldsymbol{X}_i;\theta)\}.$$

The term I_2 can be written as

$$\begin{aligned}
I_2 &= \frac{1}{n_1}\sum_{i=1}^{n} D_i \int \mathbf{1}\{s_j(\boldsymbol{X}_i;\theta_0) \le u \le s_{j+1}(\boldsymbol{X}_i;\theta_0)\} d(\widehat{F} - F_0)(u) \\
&= \int \frac{1}{n_1}\sum_{i=1}^{n} D_i \mathbf{1}\{s_j(\boldsymbol{X}_i;\theta_0) \le u \le s_{j+1}(\boldsymbol{X}_i;\theta_0)\} d(\widehat{F} - F_0)(u) \\
&= \int \widehat{c}_j(u) d(\widehat{F} - F_0)(u) \\
&= \int c_j(u) d(\widehat{F} - F_0)(u) + \int [\widehat{c}_j(u) - c_j(u)] d(\widehat{F} - F_0)(u) \\
&\equiv I_2^{(a)} + I_2^{(b)}
\end{aligned}$$

where the second line exchanges the order of summation and integration. Now, consider the term $I_2^{(b)}$. We can write it as the sum of two terms:

$$\begin{aligned}
\sqrt{n_1} I_2^{(b)} &= \sqrt{n_1}\{\mathbb{P}_{n_1}(\widehat{F}(s_m(\boldsymbol{X};\theta_0)) - F_0(s_m(\boldsymbol{X};\theta_0))) \; - P(\widehat{F}(s_m(\boldsymbol{X};\theta_0)) - F_0(s_m(\boldsymbol{X};\theta_0)))\} \\
&= \mathbb{G}_{n_1}(\widehat{F}(s_m(\boldsymbol{X};\theta_0)) - F_0(s_m(\boldsymbol{X};\theta_0))).
\end{aligned}$$

Lemma 19.24 of van der Vaart (2000) establishes that $\mathbb{G}_m(\widehat{f}_m - f_0) = o_p(1)$ for any $m \to \infty$, a random function $\widehat{f}_m$ that converges in quadratic mean to f_0, and a P-Donsker function class containing $\{\widehat{f}_m\}$. We apply this with $m = n_1$. This condition is satisfied under the current setup because

$$\int (\widehat{F}(u) - F_0(u))^2 dP(u) \le \left(\sup_{u \in \mathcal{U}} |\widehat{F}(u) - F_0(u)|\right)^2 = o_p(1).$$

where the last line follows from Proposition A.6.

Therefore, we have shown that

$$I_2 = \int c_j(u) d(\widehat{F} - F_0)(u) + o_p(n^{-1/2}).$$

The cross term I_3 measures the interaction between estimating θ and estimating F. Lemma A.11 shows that $I_3 = o_p(n^{-1/2})$. Combining the three terms yields the stated expansion. □

F 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 $\pi^{(t)}_{j|d} = \Pr(Y_{it}(0) = j \mid D_i = d)$ 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:

$$\begin{aligned}
\left(\pi^{(0)}_{j=0|d=1}, \pi^{(0)}_{j=1|d=1}, \pi^{(0)}_{j=2|d=1}\right) &= (0.3, 0.5, 0.2) \\
\left(\pi^{(1)}_{j=0|d=1}, \pi^{(1)}_{j=1|d=1}, \pi^{(1)}_{j=2|d=1}\right) &= (0.2, 0.5, 0.3) \\
\left(\pi^{(0)}_{j=0|d=0}, \pi^{(0)}_{j=1|d=0}, \pi^{(0)}_{j=2|d=0}\right) &= (0.2, 0.5, 0.3) \\
\left(\pi^{(1)}_{j=0|d=0}, \pi^{(1)}_{j=1|d=0}, \pi^{(1)}_{j=2|d=0}\right) &= (0.2, 0.4, 0.4)
\end{aligned}$$

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

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

Thus, this example demonstrates 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.

G Supplemental Information for the Empirical Application

G.1 Additional details on the application

Table 2 summarizes the methods used in the original papers.

Table 2: 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 details 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 G.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.

G.2 Analysis with covariates

I apply the semiparametric estimator from Section 4.1 to re-estimate the treatment effects conditional on pre-treatment covariates. All covariates are measured in the 2010 wave and therefore unaffected by treatment.

I construct two continuous ideology scores from a two-dimensional graded response model (GRM; Samejima, 1969) fit to 20 policy preference items in the 2010 CCES. The GRM models the ordinal response categories directly through cumulative threshold parameters, preserving the ordering information that would be lost by binarizing the items. The two-dimensional latent space is identified by fixing the mean to zero and the covariance to the identity matrix; the remaining rotational indeterminacy is resolved by varimax rotation of the factor loading matrix. The items span foreign policy (`CC304`, `CC305`), domestic policy (`CC321`, `CC324`–`CC329`), immigration (`CC322_1`–`CC322_3`), and Congressional bill approval (`CC330A`–`CC330G`, `CC330I`). The gun control item (`CC320`)

is excluded since it is the outcome variable. The model is estimated using the `mirt` package in R (Chalmers, 2012) with marginal maximum likelihood via the EM algorithm. I apply varimax rotation to the factor loading matrix and extract expected a posteriori (EAP) person scores. I assign $\beta_1 = 1$ to the more correlated dimension for normalization (the first dimension correlates at $r = 0.48$ with gun control and the second at $r = -0.45$). The demographic covariates are a female indicator, ordinal education (six levels), and consolidated race dummies. After dropping observations with missing covariates, the sample has $n = 16{,}539$ individuals.

Table 3 reports the results from the semiparametric NPMLE and the ordered probit with covariates. All confidence intervals use $B = 5{,}000$ bootstrap draws clustered at the zip code level. Both estimators find significant negative effects on category 0 (`less-strict`) under both treatment definitions. The semiparametric estimates are larger in magnitude ($\widehat{\zeta}_0 = -0.053$ for 25 miles, -0.056 for 100 miles) than the probit estimates (-0.030 and -0.036). The corresponding shifts toward categories 1 and 2 are positive but not always significant. The estimated bound of the relative effects always cover zero.

Table 3: Estimated treatment effects with covariates. Category-specific effects $\widehat{\zeta}_j$ and bounds on the relative effect $[\widehat{\tau}_L, \widehat{\tau}_U]$. Bootstrap 95% confidence intervals in parentheses ($B = 5{,}000$, clustered at the zip code level).

	25 miles	100 miles
A. Semiparametric NPMLE		
$\widehat{\zeta}_0$	-0.053	-0.056
	$(-0.071,\ -0.033)$	$(-0.069,\ -0.044)$
$\widehat{\zeta}_1$	0.024	0.032
	$(-0.003,\ 0.052)$	$(0.015,\ 0.051)$
$\widehat{\zeta}_2$	0.028	0.024
	$(0.004,\ 0.049)$	$(0.007,\ 0.038)$
$[\widehat{\tau}_L,\ \widehat{\tau}_U]$	$[-0.104,\ 0.213]$	$[-0.135,\ 0.240]$
	$(-0.128,\ 0.240)$	$(-0.151,\ 0.258)$
B. Ordered probit with covariates		
$\widehat{\zeta}_0$	-0.030	-0.036
	$(-0.045,\ -0.015)$	$(-0.046,\ -0.026)$
$\widehat{\zeta}_1$	0.012	0.021
	$(-0.012,\ 0.035)$	$(0.007,\ 0.036)$
$\widehat{\zeta}_2$	0.018	0.015
	$(-0.002,\ 0.038)$	$(0.002,\ 0.027)$
$[\widehat{\tau}_L,\ \widehat{\tau}_U]$	$[-0.115,\ 0.180]$	$[-0.145,\ 0.210]$
	$(-0.137,\ 0.202)$	$(-0.159,\ 0.225)$

G.3 Analysis by prior exposure to mass shootings

I stratify the two-wave analysis by whether respondents experienced mass shootings within 100 miles in the decade before 2010 ("prior exposure" versus "no prior exposure"). Figure G.2 presents

the results. The category-specific effects and the bounds on the relative effect are close to zero in both subgroups, and all 95% confidence intervals cover zero. The pattern is consistent with the pooled results in Figure 2.

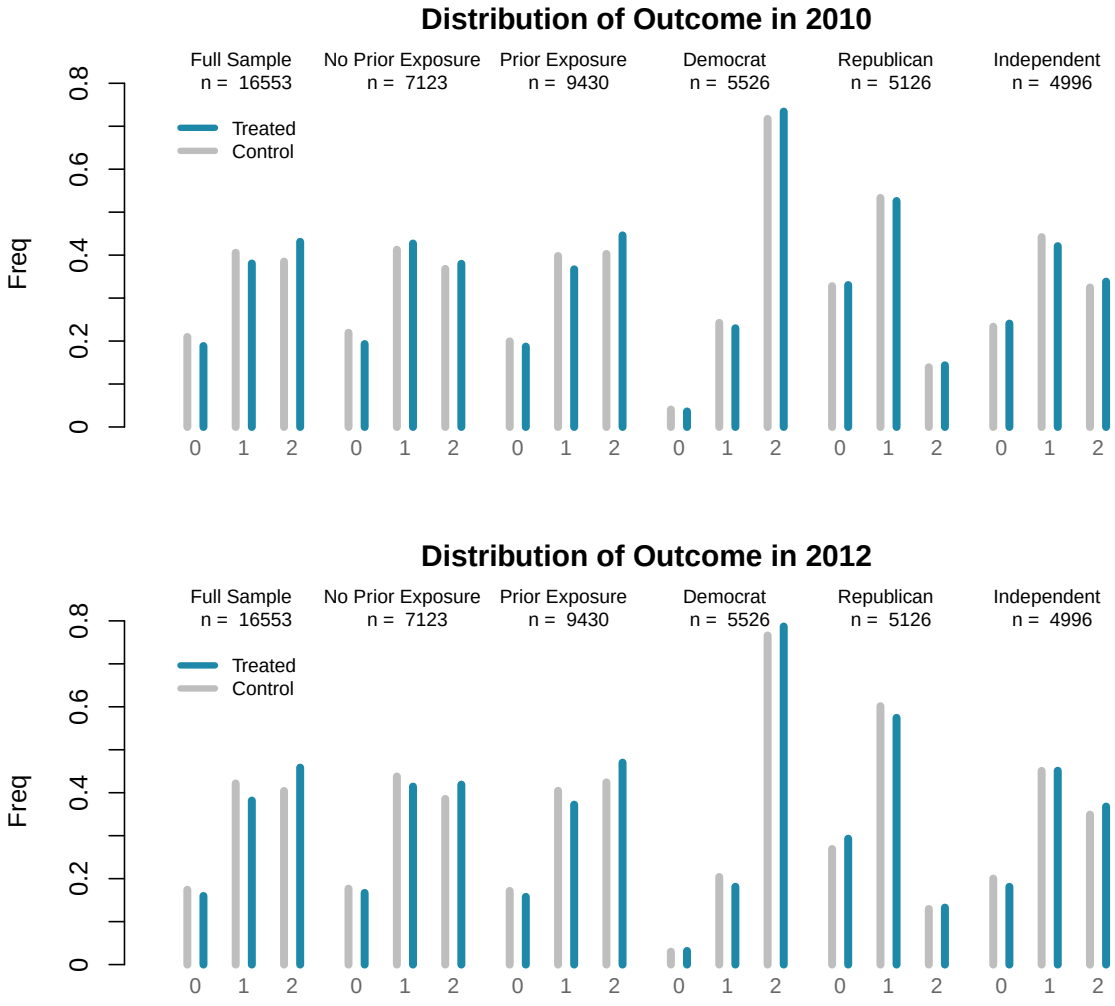

Figure G.1: Distribution of outcomes: `(0): less-strict`, `(1): kept-as-they-are` and `(2): more-strict`. The top panel shows the distribution of 2010 and the bottom panel is for 2012. Bars in blue (gray) show distributions for the treated (control) group.

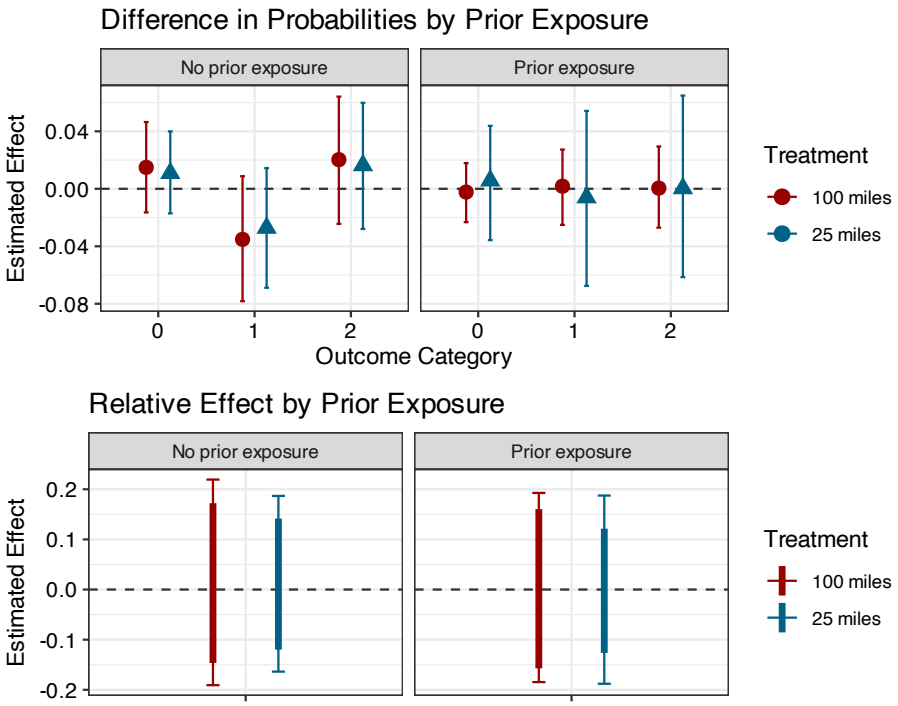

Figure G.2: Upper panels – Estimated treatment effects $\widehat{\zeta}_j$ by prior exposure to mass shootings with 95% confidence intervals. Circles indicate effects under the 100 mile threshold, while triangles indicate effects under the 25 mile threshold. **Lower panels** – Estimated bounds on the relative effect $[\widehat{\tau}_L, \widehat{\tau}_U]$ (thick lines) with 95% confidence intervals (thin lines).