When Is Parallel Trends Sensitive to Functional Form?*

Jonathan Roth† Pedro H.C. Sant'Anna‡

February 1, 2021

Abstract

This paper assesses when the validity of difference-in-differences and related estimators depends on functional form. We provide a novel characterization: the parallel trends assumption holds under all strictly monotonic transformations of the outcome if and only if a stronger "parallel trends"-type condition holds for the cumulative distribution function of untreated potential outcomes. This condition is satisfied if and essentially only if the population can be partitioned into a subgroup for which treatment is effectively randomly assigned and a remaining subgroup for which the distribution of untreated potential outcomes is stable over time. We show further that it is impossible to construct any estimator that is consistent (or unbiased) for the average treatment effect on the treated (ATT) without either imposing functional form restrictions or imposing assumptions that identify the full distribution of untreated potential outcomes. Our results suggest that researchers who wish to point-identify the ATT should justify one of the following: (i) why treatment is as-if randomly assigned, (ii) why the chosen functional form is correct at the exclusion of others, or (iii) a method for inferring the entire counterfactual distribution of untreated potential outcomes.

*We thank Isaiah Andrews, Otavio Bartalotti, Kirill Borusyak, Kevin Chen, Carol Caetano, Arin Dube, Dalia Ghanem, Andrew Goodman-Bacon, Ryan Hill, Martin Huber, Peter Hull, Ariella Kahn-Lang, Kevin Lang, Arthur Lewbel, Daniel Millimet, Aureo de Paula, Ashesh Rambachan, Adrienne Sabety, Yuya Sasaki, Jesse Shapiro, Tymon Słoczyński, Alex Torgovitsky, Kaspar Wüthrich, and seminar participants at Brandeis, UC-Davis, the Southern Economic Association annual conference, and the Chamberlain Seminar for helpful comments and conversations.

†Microsoft. Email: Jonathan.Roth@microsoft.com

‡Vanderbilt University. Email: pedro.h.santanna@vanderbilt.edu

1 Introduction

Difference-in-differences (DiD) is one of the most popular strategies in the social sciences for estimating causal effects in non-experimental contexts. The DiD design allows for identification of the average treatment effect on the treated (ATT) under the so-called "parallel trends" assumption, which is weaker than that of (as-if) random assignment of treatment. This paper studies the content of the parallel trends assumption in settings where treatment may not be (as-if) randomly assigned.

We focus on the extent to which the assumptions underlying DiD and other estimators of the ATT depend on the functional form of the outcome. Following Athey and Imbens (2006), we say that an assumption is robust to functional form if it is invariant to strictly monotonic transformations of the outcome — i.e., if the assumption holds for potential outcomes $Y(\bullet)$, then it also holds if the potential outcomes are replaced with any strictly monotonic function of the original potential outcomes $g(Y(\bullet))$.[1] Intuitively, this invariance requires that the validity of the assumption does not depend on the units in which the outcome is measured.

The motivation for studying this property is that it is often not obvious from theory which is the "right" transformation of the outcome for an identifying assumption such as parallel trends to hold. As an example, various studies of labor market earnings have used as the outcome earnings in levels, earnings in logs, the inverse hyperbolic sine of earnings, or the percentile of earnings in the national distribution. These are non-linear and sometimes discontinuous transformations of the same outcome.[2] If we are considering a difference-in-differences design for earnings, which (if any) of these transformations is appropriate? Economic theory will often not be informative as to which is the right transformation for parallel trends to hold, and so it will often be desirable if the validity of the research design does not depend on this choice. Invariance to transformations will also be desirable in settings where the outcome does not have natural units, such as test scores (Cunha and Heckman, 2008).

Indeed, concerns about sensitivity to functional form have a long history in econometrics. For instance, Leamer's (1983; 1985) influential critiques of applied econometrics in the 1980s focused heavily on functional form considerations. As described in Angrist and Pischke (2010), such "functional form concerns [have become] less central" more recently in light

[1] Technically, we need to restrict attention to measurable functions g for which the expectation is finite. Athey and Imbens (2006, FN 11) use the phrase "invariance to scale" to describe this property; we use "invariance to transformations" instead to make clear that the transformations may be non-linear.

[2] Percentiles of the earnings distribution will sometimes be discontinuous functions of earnings, since the earnings distribution exhibits bunching at round numbers and tax thresholds.

of a renewed focus on research design. This is because under design-based assumptions about a how treatment (or instrument) is assigned, ordinarily least squares (OLS) or linear instrumental variables (IV) recover a weighted average of causal effects even if the linear model is mis-specified. However, the usual starting point for difference-in-differences is the parallel trends assumption, which does not directly impose restrictions on the treatment assignment mechanism. In this paper, we therefore explore to what extent functional form is a concern for difference-in-differences and related estimators.

Our first main result characterizes when the parallel trends assumption is invariant to transformations. We prove that the parallel trends assumption is invariant to transformations if and only if a "parallel trends"-type condition holds for the entire cumulative distribution function (CDF) of untreated potential outcomes. There are three cases in which this condition holds: first, if the distribution of potential outcomes is the same for both groups, as occurs under random assignment of treatment. Second, if the potential outcome distributions for each group are stable over time. And third, a hybrid of the first two cases in which the population is essentially a mixture of a sub-population that is effectively randomized between treatment and control and another sub-population that has non-random treatment status but stable potential outcome distributions across time. In settings where the treatment is not (as-if) randomly assigned, the assumptions needed for the invariance to transformations of parallel trends will thus often be quite restrictive.

Our second main result shows that the ATT is identified under all strictly monotonic transformations of the outcome if and only if the entire distribution of untreated potential outcomes for the treated group is identified. Thus, to obtain *any* consistent estimator of the ATT, one must impose assumptions that either are sensitive to functional form *or* that identify the full counterfactual distribution of potential outcomes. Depending on context, however, different methods for inferring the counterfactual distribution may be preferable to others. Indeed, the changes-in-changes (CiC) model of Athey and Imbens (2006), the distributional DiD models of Bonhomme and Sauder (2011) and Callaway and Li (2019), and the condition described above for parallel trends to be invariant to transformations all suggest different, non-nested ways of imputing the distribution of counterfactual outcomes. If one wishes to obtain a consistent estimator without imposing assumptions that depend on functional form, it is thus natural to first choose a method for inferring the counterfactual distribution, and then choose an estimator that is consistent under those assumptions. Our identification result also implies that it is no accident that the CiC model both identifies the ATT without imposing functional form restrictions and identifies quantile treatment effects: in order to obtain such invariance to functional form, it is *necessary* to take a stand on how to infer the entire counterfactual distribution.

Our results have important implications for practitioners considering the use of a difference-in-differences design. In light of our results, researchers interested in point-identifying the ATT should make one of the following three justifications. First, they may argue that treatment is (as-if) randomly assigned. In this case, parallel trends will hold under all transformations of the outcome.[3] Second, the researcher may argue for a particular method of inferring the counterfactual distribution of potential outcomes for the treated group, and choose an appropriate estimator that is valid regardless of the functional form under this assumption. Third, the researcher may give up on robustness to transformations, and argue for the validity of the particular chosen functional form at the (necessary) exclusion of others. If none of these justifications is appealing, the researcher may instead impose weaker assumptions on the data-generating process that do not point-identify the ATT, e.g. using partial identification tools that do not impose that the parallel trends assumption holds exactly (Manski and Pepper, 2018; Rambachan and Roth, 2020b).

Several previous papers have noted that the parallel trends assumption may be sensitive to functional form, with particular attention paid to the logs versus levels specifications (Meyer, 1995; Athey and Imbens, 2006; Lechner, 2011; Kahn-Lang and Lang, 2020; Ding and Li, 2019). To our knowledge, however, we are the first to provide a complete characterization for the invariance to transformations of the parallel trends assumption. Moreover, we show that a widely-held intuition about the sensitivity of the parallel trends assumption is not quite correct. Specifically, it has been previously stated that the parallel trends assumption in levels is incompatible with the parallel trends assumption in logs if baseline distributions differ between the treatment and comparison groups (Meyer, 1995; Angrist and Pischke, 2009; Kahn-Lang and Lang, 2020). Although it is true that parallel trends may hold in logs but not levels (or vice versa), we show that having identical baseline distributions is not necessary (nor sufficient) for the parallel trends assumption to be invariant to transformations.

Our work relates to several papers that consider identification of quantile treatment effects in difference-in-differences settings (Athey and Imbens, 2006; Bonhomme and Sauder, 2011; Callaway and Li, 2019). These papers introduce new sets of assumptions that differ from the usual parallel trends assumption and allow for identification of the full distribution of untreated potential outcomes. One feature of the set of assumptions introduced in Athey and Imbens (2006) is that it is invariant to transformations of the outcome, unlike the usual parallel trends assumption. By contrast, we derive conditions on the distributions of

[3] We note, however, that other estimators, such as the simple difference-in-means and the analysis of covariance (ANCOVA), will also be consistent for the ATT under randomization of treatment and may be preferred for efficiency reasons (McKenzie, 2012).

potential outcomes under which the usual parallel trends assumption is invariant to transformations. These conditions are different from, and non-nested with, the conditions provided for identification of the full distribution of potential outcomes in previous work.

Our results also draw connections between approaches to causal identification that rely on an unconfoundedness assumption (e.g. Imbens and Rubin (2015)) and difference-in-differences designs, which rely on the parallel trends assumption. We show, roughly speaking, that parallel trends is either a functional form restriction or a combination of unconfoundedness and stationarity assumptions.

We are not aware of any previous papers showing the simple but important fact that identification of the ATT under all monotonic transformations is equivalent to identification of the distribution of counterfactual outcomes for the treated group. A powerful implication of this result is that the consistency of any estimator for the ATT must either depend on functional form assumptions or assumptions that identify the full distribution of untreated potential outcomes. This result does not depend on the multi-period structure of DiD, and may be relevant in other contexts where ATTs are of interest.

2 Model

We consider a canonical two-period difference-in-differences model. There are two periods $t = 0, 1$, and units indexed by i come from one of two populations denoted by $D_i \in \{0, 1\}$. Units in the $D_i = 1$ (treated) population receive treatment beginning in period $t = 1$, and units in the $D_i = 0$ (comparison) population never receive treatment. We denote by $Y_{i,t}(1), Y_{i,t}(0)$ the potential outcomes for unit i in period t under treatment and control, respectively, and we observe the outcome $Y_{i,t} = D_i Y_{i,t}(1) + (1 - D_i)Y_{i,t}(0)$, where D_i is an indicator for whether unit i is in the treated or comparison population. We assume that there are no anticipatory effects of treatment, so that $Y_{i,t=0}(1) = Y_{i,t=0}(0)$ for all i.[4] The average treatment effect on the treated is defined as

$$\tau_{ATT} = \mathbb{E}\left[Y_{i,t=1}(1) - Y_{i,t=1}(0) \,|\, D_i = 1\right].$$

Remark 1 (Multiple periods and staggered timing)**.** We consider here a two period model for expositional simplicity. Several recent papers have considered settings with multiple periods and staggered treatment timing under a generalized parallel trends assumption that imposes the two-period, two-group version of parallel trends for multiple pairs of cohorts and periods (e.g., Assumptions 4 and 5 in Callaway and Sant'Anna (2020), Assumption 1 in

[4] This assumption may be violated if units adjust behavior in anticipation of treatment (Malani and Reif, 2015).

Sun and Abraham (2020), or Assumption 5 in de Chaisemartin and D'Haultfœuille (2020)). Our results on the parallel trends assumption in this simple model thus have immediate implications for the generalized parallel trends assumption in the staggered case.

Remark 2 (Conditional parallel trends)**.** Likewise, for simplicity we consider a model that does not condition on unit-specific covariates. However, the same results would go through if all probability statements were conditional on some value of unit-specific covariates X_i. Our results thus have implications for the conditional parallel trends assumptions considered in Abadie (2005); Heckman, Ichimura and Todd (1997); Callaway and Sant'Anna (2020); Sant'Anna and Zhao (2020).

Remark 3 (Sampling-based versus design-based uncertainty)**.** In the main text of the paper, we adopt a sampling-based (also known as model-based) view of uncertainty, which facilitates comparison to previous work on DiD. In some contexts, however, the view of sampling from a super-population may be unnatural, e.g. when we observe outcomes for all 50 U.S. states (Manski and Pepper, 2018). Recent papers by Athey and Imbens (2018) and Rambachan and Roth (2020a) have studied DiD from a design-based view of uncertainty in which the population is treated as fixed and the random nature of the data arises from the stochastic assignment of treatment. In the Appendix, we derive similar results on the invariance to transformations of the DiD estimator from a design-based perspective.

3 Invariance of Parallel Trends

The classical assumption that allows for point identification of the ATT in the DiD design is the so-called parallel trends assumption, which imposes that

$$\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 0\right]. \tag{1}$$

Under the parallel trends assumption, $\tau_{ATT} = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$, where $\mu_{ds} = \mathbb{E}\left[Y_{i,t=s} \mid D_i = d\right]$. We assume throughout that the four expectations in (1) exist and are finite. Following Athey and Imbens (2006), we say that the parallel trends assumption is invariant to transformations if the parallel trends assumption holds for all strictly monotonic transformations of the outcome.

Definition 1. We say that the parallel trends assumption is invariant to transformations if

$$\begin{aligned}\mathbb{E}\left[g(Y_{i,t=1}(0)) \mid D_i = 1\right] - \mathbb{E}\left[g(Y_{i,t=0}(0)) \mid D_i = 1\right]& \\ = \mathbb{E}\left[g(Y_{i,t=1}(0)) \mid D_i = 0\right] - \mathbb{E}\left[g(Y_{i,t=0}(0)) \mid D_i = 0\right]&\end{aligned}$$

for all strictly monotonic, measurable functions g such that the expectations above are finite.

Our first main result characterizes when parallel trends is invariant to transformations.

Proposition 3.1. *Parallel trends is invariant to transformations if and only if*

$$F^{Y(0)}_{D=1,t=1}(y) - F^{Y(0)}_{D=1,t=0}(y) = F^{Y(0)}_{D=0,t=1}(y) - F^{Y(0)}_{D=0,t=0}(y), \text{ for all } y \in \mathbb{R} \tag{2}$$

where $F^{Y(0)}_{D=d,t=s}$ is the cumulative distribution function of $Y_{i,t=s}(0) \mid D_i = d$.

Proof. If (2) holds, then from integrating on both sides of the equation it is immediate that

$$\int g(y)dF^{Y(0)}_{D=1,t=1} - \int g(y)dF^{Y(0)}_{D=1,t=0} = \int g(y)dF^{Y(0)}_{D=0,t=1} - \int g(y)dF^{Y(0)}_{D=0,t=0} \tag{3}$$

for any strictly monotonic measurable g such that the integrals exist and are finite, and hence parallel trends is invariant to transformations.

Conversely, if parallel trends is invariant to transformations, then (3) holds for every strictly monotonic, measurable g such that the expectations exist and are finite. In particular, it holds for the identity map $g_1(y) = y$ as well as the map $g_2(y) = y - 1[y \leqslant \tilde{y}]$ for any given $\tilde{y} \in \mathbb{R}$. Then, it follows that

$$\begin{aligned}
&\int ydF^{Y(0)}_{D=1,t=1} - \int ydF^{Y(0)}_{D=1,t=0} = \int ydF^{Y(0)}_{D=0,t=1} - \int ydF^{Y(0)}_{D=0,t=0}, \text{ and} \\
&\int (y - 1[y \leqslant \tilde{y}])dF^{Y(0)}_{D=1,t=1} - \int (y - 1[y \leqslant \tilde{y}])dF^{Y(0)}_{D=1,t=0} = \\
&\int (y - 1[y \leqslant \tilde{y}])dF^{Y(0)}_{D=0,t=1} - \int (y - 1[y \leqslant \tilde{y}])dF^{Y(0)}_{D=0,t=0}.
\end{aligned}$$

Subtracting the second equation from the first in the previous display, we obtain

$$\int 1[y \leqslant \tilde{y}]dF^{Y(0)}_{D=1,t=1} - \int 1[y \leqslant \tilde{y}]dF^{Y(0)}_{D=1,t=0} = \int 1[y \leqslant \tilde{y}]dF^{Y(0)}_{D=0,t=1} - \int 1[y \leqslant \tilde{y}]dF^{Y(0)}_{D=0,t=0},$$

which is equivalent to (2) by the definition of the CDF and the fact that $\tilde{y}$ is arbitrary. ☐

Proposition 3.1 shows that parallel trends is invariant to transformations if and only if a "parallel trends"-type assumption holds for the CDFs of the untreated potential outcomes. We note that if the outcome is continuous, then parallel trends of CDFs is equivalent to parallel trends of PDFs (almost everywhere). The following result provides a characterization of how distributions satisfying this assumption can be generated.

Proposition 3.2. *Suppose that the distributions $Y_{i,t=s}(0)|D_i = d$ for all $d, s \in \{0, 1\}$ have* *a Radon-Nikodym density with respect to a common dominating, positive σ-finite measure.*[5] *Then condition (2) holds if and only if*

$$F_{D=d,t=s}^{Y(0)}(y) = \theta F_{t=s}^{Y(0)}(y) + (1-\theta) F_{D=d}^{Y(0)}(y) \text{ for all } y \in \mathbb{R} \text{ and } d, s \in \{0,1\}, \tag{4}$$

where $\theta \in [0,1]$ and $F_{t=s}^{Y(0)}(y)$ and $F_{D=d}^{Y(0)}(y)$ are CDFs of distributions that depend only on time and group, respectively.

Proof. See Appendix A.1. □

Proposition 3.2 shows that parallel trends of CDFs is satisfied if and only if the untreated potential outcomes for each group and time can be represented as a mixture of a common time-varying distribution that does not depend on group (with weight θ) and a group-specific distribution that does not depend on time (with weight $1-\theta$). There are three cases in which this will be satisfied, depending on the value of θ.

Case 1: Random assignment. ($\boldsymbol{\theta = 1}$). The case $\theta = 1$ corresponds with imposing that the distributions of $Y(0)$ for the treated and comparison groups are the same in each period, $F_{D=1,t}^{Y(0)}(y) = F_{D=0,t}^{Y(0)}(y)$, as occurs under (as-if) random assignment of treatment.

Case 2: Stationary $\boldsymbol{Y(0)}$. ($\boldsymbol{\theta = 0}$). The case $\theta = 0$ corresponds with imposing that the distribution of $Y(0)$ for both the treated and comparison populations does not depend on time, i.e. $F_{D=d,t=1}^{Y(0)}(y) = F_{D=d,t=0}^{Y(0)}(y)$.

Case 3: Non-random assignment and non-stationarity. ($\boldsymbol{\theta \in (0,1)}$). Remarkably, if (4) holds with $\theta \in (0,1)$, then parallel trends is invariant to transformations even though there are both differences in distributions between the treated and comparison groups and non-stationary potential outcomes.[6] The case $\theta \in (0,1)$ corresponds with a hybrid of the first two cases. Intuitively, it captures the situation in which the population is composed of three types of units: type A is effectively randomized between the treated and comparison populations, composing θ fraction of each population, whereas Types B and C compose the remainder of the treated and comparison groups, respectively, and have stable distributions of potential outcomes over time.[7] Although quite specific, this case might be plausible if treatment is

[5] This condition is satisfied if $Y_{i,t=s}(0)|D=d$ is continuously distributed (using the probability density function and Lebesgue measure), or discrete with finite support (using the probability mass function and counting measure). The condition will also be satisfied for many non-pathological mixed distributions.

[6] This assumes that $F_{t=1}^{Y(0)}(y) \neq F_{t=0}^{Y(0)}(y)$ for some y and $F_{D=1}^{Y(0)}(y) \neq F_{D=0}^{Y(0)}(y)$ for some y. Otherwise, the CDFs can be re-written such that (4) holds with either $\theta = 0$ or $\theta = 1$.

[7] Proposition 3.2 implies that θ fraction of the treated and control populations have a common distribution in each period, and the remaining $1-\theta$ fraction of each group has a group-specific distribution that does not

as-if randomly assigned among a subset of the population for which we expect there to be time trends (e.g., younger workers on an upward earnings trajectory), but there are also units in the population for whom treatment status is endogenous but we expect outcomes to be relatively stable over time (e.g. older workers with stable earnings trajectory).[8] This is perhaps the most interesting case, since in the other two cases a single difference (either across time or across groups) would suffice to identify the ATT.

We now provide several remarks that further clarify and contextualize our results.

Remark 4 (Empirical papers using parallel trends of CDFs)**.** Several empirical papers have used a DiD design to estimate the effects of a treatment on the entire CDF (or density) of an outcome (Almond, Hoynes and Schanzenbach, 2011; Dube, 2019; Stepner, 2019). For instance, Cengiz, Dube, Lindner and Zipperer (2019) and Dube and Lindner (2020) analyze the effects of changing the minimum wage on the share of the population earning any given wage or more.[9] Proposition 3.1 implies that the parallel trends assumption underlying the validity of these distributional analyses is equivalent to the invariance to transformations of the usual parallel trends assumption needed to identify the ATT. Future work using DiD to identify distributional effects might use the three cases implied by Proposition 3.2 to justify the necessary parallel trends of distributions.

Remark 5 (Binary outcomes)**.** Suppose the outcome is binary, $Y_i \in \{0,1\}$. Then for any $y \in [0,1)$, $F^{Y(0)}_{D=1,t=1}(y) = 1 - \mathbb{E}[Y_{i,t=1}(0) \mid D_i = 1]$, and analogously for the other CDFs. Thus, (2) is equivalent to the parallel trends assumption (1). Proposition 3.1 thus implies that whenever the parallel trends assumption holds, it also holds for all monotonic transformations of the outcome. This is intuitive, as the expectation of a binary outcome fully characterizes its distribution. Note that this does not imply that the parallel trends assumption necessarily holds for binary outcomes, only that it does not depend on the transformation of the outcome.

Remark 6 (Restrictiveness in some settings)**.** In settings where the treatment is not randomly assigned and the outcome is not binary, the condition in (2) will generally be stronger than parallel trends and may often be restrictive. Indeed, note that (2) is equivalent to

$$F^{Y(0)}_{D=1,t=1}(y) = F^{Y(0)}_{D=1,t=0}(y) + F^{Y(0)}_{D=0,t=1}(y) - F^{Y(0)}_{D=0,t=0}(y). \tag{5}$$

depend on time. In principle, it is possible for the units in the θ and $1-\theta$ sub-populations to change over time, although it is difficult to imagine scenarios where this would be the case. We thus write that parallel trends can be invariant to transformations "essentially only if" the population can be partitioned into groups such that one is effectively randomized, and the others have stable outcomes over time.

[8]Of course, if data on worker age were available, then the researcher could restrict attention to younger workers, in which case we would be back in Case 1.

[9]In practice, the wages are discretized and normalized relative to the minimum wage in a reference year.

The left-hand side of (5) is a CDF and therefore must be non-decreasing and bounded between $[0, 1]$, but this is not guaranteed for the right-hand side. To further highlight the potential restrictiveness of this condition, we show in Appendix A.2 that if $\mu_1, ..., \mu_4$ are distinct, then it is impossible to have

$$\Phi_{\mu_1,\sigma_1}(y) = \Phi_{\mu_2,\sigma_2}(y) + \Phi_{\mu_3,\sigma_3}(y) - \Phi_{\mu_4,\sigma_4}(y), \tag{6}$$

for all $y \in \mathbb{R}$, where $\Phi_{\mu,\sigma}$ is the CDF of the $\mathcal{N}(\mu, \sigma^2)$ distribution and the σ_j^2 are arbitrary positive variances (possibly non-distinct). Thus, with normally distributed outcomes, parallel trends will be sensitive to functional form unless either treatment is as-if randomly assigned or $Y(0)$ is stationary (Cases 1 and 2 described above).

Remark 7 (Falsifiability of the invariance condition)**.** As mentioned, condition (5) cannot be satisfied if the terms on the right-hand side, which are CDFs of identified distributions, violate the properties required of a CDF. The invariance to transformations of parallel trends thus has falsifiable implications. In fact, it is straightforward to verify that a sharp testable implication is that the right-hand side of (5) is monotonic.[10] One can therefore use off-the-shelf methods to construct a falsification test based on the null hypothesis of monotonicity. For instance, methods like those developed in Delgado and Escanciano (2013) and Fang (2019) could be used to test whether the integrated curve

$$C(y) = -\int_{-\infty}^{y} (F_{D=1,t=0}^{Y(0)}(u) + F_{D=0,t=1}^{Y(0)}(u) - F_{D=0,t=0}^{Y(0)}(u))dG(u)$$

is concave, where $G(\cdot)$ is the CDF of the pooled outcome among observations with $(D = 1, t = 0)$, $(D = 0, t = 1)$, $(D = 0, t = 0)$. Alternatively, one could test whether the continuum of moment inequalities

$$\mathbb{E}\left[w\left(D, T\right) 1\left\{y' < Y \leqslant y\right\}\right] \geqslant 0 \text{ for all } (y, y') \in \mathbb{R}^2 \;:\; y > y',$$

is satisfied, e.g. using the methods in Andrews and Shi (2013), where

$$w\left(D, T\right) = \frac{D\left(1-T\right)}{\mathbb{E}\left[D\left(1-T\right)\right]} + \frac{\left(1-D\right)T}{\mathbb{E}\left[\left(1-D\right)T\right]} - \frac{\left(1-D\right)\left(1-T\right)}{\mathbb{E}\left[\left(1-D\right)\left(1-T\right)\right]},$$

Y is the pooled outcome data from all groups, and T and D are dummy variables for the outcome being from time $t = 1$ and from the treated group, respectively.

[10]It follows immediately from the properties of CDFs that the RHS of (5) is right-continuous and has limits of 0 and 1 as $y \to -\infty$ and $y \to \infty$, respectively. Appendix C provides an alternative characterization of when the condition for invariance will be falsified. This occurs if and only if the identified distributions of $Y(0) \mid D = d, t = s$ are sufficiently far apart (in total variation distance) in both the time and group dimensions.

We urge caution in relying too much on such falsification tests, however. First, such tests are for the null that there is *some* possible counterfactual distribution for the treated group in period 1 such that parallel trends is invariant to transformations. Yet even if the null is satisfied there is no guarantee that the distribution that leads to invariance would have been realized under the counterfactual. In other words, parallel trends of CDFs is *falsifiable* but *not verifiable*. Second, as with tests of pre-existing trends (Roth, 2020), there are concerns that such tests may have low power against violations of the null in finite samples, and conditioning the analysis on the result of such a pre-test may distort estimation and inference. We therefore encourage researchers to focus on *ex ante* justifications for the parallel trends assumption.

Remark 8 (Relationship to Athey and Imbens (2006)'s changes-in-changes model)**.** The condition in equation (2) needed for parallel trends to be invariant to transformations may be reminiscent of the "changes-in-changes" (CiC) model of Athey and Imbens (2006). The two are not equivalent, however. As noted above, (2) implies that

$$F^{Y(0)}_{D=1,t=1}(y) = F^{Y(0)}_{D=1,t=0}(y) + F^{Y(0)}_{D=0,t=1}(y) - F^{Y(0)}_{D=0,t=0}(y), \tag{7}$$

whereas the Athey and Imbens (2006) model implies that

$$F^{Y(0)}_{D=1,t=1}(y) = F^{Y(0)}_{D=1,t=0}(F^{Y(0),-1}_{D=0,t=0}(F^{Y(0)}_{D=0,t=1}(y))), \tag{8}$$

where $F^{Y(0),-1}_{D=0,t=0}(\tau) = \inf\{x \in \mathbb{R} : F^{Y(0)}_{D=0,t=0}(x) \geqslant \tau\}$ is the τ-quantile of $Y_{i,t=0}(0)$ among untreated units. Both equations are satisfied under random assignment of treatment or stationary potential outcomes (Cases 1 and 2), in which case the right-hand side of equations (7) and (8) both reduce to $F^{Y(0)}_{D=0,t=1}(y)$ or $F^{Y(0)}_{D=1,t=0}(y)$, respectively. Outside of these cases, however, the two conditions are non-nested. For instance, equation (8) will not generally hold in Case 3 given above (except for special choices of the distributions), since the mapping between quantiles for the treated and untreated groups need not be preserved across periods.[11] Conversely, recall that parallel trends of CDFs will necessarily be violated if the distributions of potential outcomes are such that the right-hand side of (5) is non-monotonic, whereas Athey and Imbens (2006) show that with continuous outcomes one can always construct a distribution for the treated group in $t = 1$ such that (8) is satisfied. Parallel trends of CDFs is also non-nested with the quantile difference-in-differences model discussed in

[11] As a concrete counterexample, suppose $\theta = 0.5$, $F^{Y(0)}_t$ is the CDF for the uniform distribution on $[0, 1]$ for $t = 0$ and the uniform distribution on $[1, 2]$ for $t = 1$, whereas $F^{Y(0)}_d$ corresponds with the CDF of a point mass at 1 and 0 for $d = 0$ and $d = 1$, respectively. Then $F^{Y(0)}_{D=1,t=1}(1.2) = 0.6$, whereas equation (8) implies it is 0.1.

Athey and Imbens (2006), which imposes a parallel-trends type assumption on the inverse CDF.

Remark 9 (Relationship to Bonhomme and Sauder (2011)'s distributional DiD model)**.** Condition (2) is also non-nested with the distributional DiD model proposed by Bonhomme and Sauder (2011). Their model implies a "parallel trends" condition for the log of the characteristic function, i.e.,

$$log\Psi^{Y(0)}_{D=1,t=1}(s) = log\Psi^{Y(0)}_{D=1,t=0}(s) + log\Psi^{Y(0)}_{D=0,t=1}(s) - log\Psi^{Y(0)}_{D=0,t=0}(s), \tag{9}$$

where, for instance, $\Psi^{Y(0)}_{D=1,t=1}(\cdot)$ is the characteristic function of $Y_{i,t=1}(0) \mid D_i = 1$. By contrast, condition (2) implies a parallel trends assumption for the levels of the characteristic function,[12]

$$\Psi^{Y(0)}_{D=1,t=1}(s) = \Psi^{Y(0)}_{D=1,t=0}(s) + \Psi^{Y(0)}_{D=0,t=1}(s) - \Psi^{Y(0)}_{D=0,t=0}(s). \tag{10}$$

Conditions (9) and (10) are both satisfied under random assignment of treatment and stationary $Y(0)$ (Cases 1 and 2), but otherwise they are generally non-nested. For instance, as with the CiC model, condition (9) will not generally hold in Case 3 above.[13] Conversely, if the potential outcomes (conditional on group and period) are normally distributed with equal variances, then (9) holds if parallel trends holds in levels, whereas we showed in Remark 6 that this is not the case for equation (2).[14]

Remark 10 (Relationship to Callaway and Li (2019)'s distributional DiD model)**.** We also note that condition (2) differs from the identifying assumptions for the distributional DiD model proposed by Callaway and Li (2019), which are: 1) that $Y_{i,t=1}(0) - Y_{i,t=0}(0)$ is fully independent of the treatment assignment D_i, and 2) a copula stability assumption, which states that the dependence between $Y_{i,t=1}(0) - Y_{i,t=0}(0)$ and $Y_{i,t=0}(0)$ among treated units is the same as the the dependence between $Y_{i,t=0}(0) - Y_{i,t=-1}(0)$ and $Y_{i,t=-1}(0)$ among treated units.[15] While the full independence assumption is guaranteed to hold under random assignment of treatment, the copula stability assumption is not guaranteed to hold even under randomization. Thus, Callaway and Li (2019)'s identifying assumptions may not coincide

[12]To see why this is the case with continuously distributed outcomes, differentiate both sides of (2) to obtain parallel trends of PDFs, then apply a Fourier transform to both sides and use the linearity of the Fourier transform to obtain parallel trends of characteristic functions.

[13]It fails for the same example as in footnote 11.

[14]This follows from the fact that the characteristic function of the $\mathcal{N}\left(\mu,\, \sigma^2\right)$ variable is $exp(\mu it - \frac{1}{2}\sigma^2 t^2)$.

[15]Note that the model of Callaway and Li (2019) relies on having access to data for three time periods, in contrast to the other models considered so far. See also Callaway, Li and Oka (2018) who consider a related set of identifying assumptions that requires data from only two time periods.

with the other distributional models discussed so far under random assignment of treatment. Nonetheless, in non-experimental settings it can be the case that Callaway and Li (2019) assumptions hold whereas (2) does not (and vice-versa).

Remark 11 (Equality of baseline distributions)**.** It has previously been stated that the parallel trends assumption in levels is incompatible with the parallel trends assumption in logs, unless the distibution of baseline outcomes are the same for the two groups. For example, Angrist and Pischke (2009, p. 230) write,

> "[C]ommon trends in logs rule out common trends in levels and vice versa"

and Kahn-Lang and Lang (2020, p. 615) write,

> "[U]nless the distribution of outcomes is initially the same for the experimental and control groups, the effect of any changes associated with time cannot be the same both if the model is specified in, for example, levels and if it is specified in logarithms."

Case 3 above makes clear that this is not quite correct, however: parallel trends can be invariant to transformation even if baseline distributions are different and there are time trends. We nonetheless agree with Kahn-Lang and Lang (2020)'s qualitative statement that the "choice of functional form ... requires justification." We also agree with Kahn-Lang and Lang (2020)'s assertion that equality of pre-treatment outcomes is not *sufficient* for parallel trends to be invariant to transformations, since the parallel trends assumption also places restrictions on counterfactual outcomes in period 1.

Remark 12 (Equality of baseline means)**.** In fact, Propositions 3.1 and 3.2 imply that parallel trends can be invariant to transformations even if the baseline means of $Y(0)$ differ for the treated and comparison groups. This will be the case if the decomposition in (4) holds with $\theta > 0$ and CDFs $F_{D=1}^{Y(0)}$ and $F_{D=0}^{Y(0)}$ that correspond with distributions with different means. This may be initially unintuitive, since if $\mu_1, ..., \mu_4$ are distinct, it cannot be the case that both $\mu_1 - \mu_2 = \mu_3 - \mu_4$ and $log(\mu_1) - log(\mu_2) = log(\mu_3) - \log(\mu_4)$, i.e. that the level differences and percentage differences between the means are the same. However, this fact only implies that if the means $\mathbb{E}\left[Y_{i,t=s}(0) \mid D_i = d\right]$ are distinct for all (d, s), then it cannot simultaneously be true that the parallel trends assumption (1) holds and that

$$\begin{aligned} &log\left(\mathbb{E}\left[Y_{i,t=1}(0)|D_i = 1\right]\right) - log\left(\mathbb{E}\left[Y_{i,t=0}(0)|D_i = 1\right]\right) \\ &\qquad = log\left(\mathbb{E}\left[Y_{i,t=1}(0)|D_i = 0\right]\right) - log\left(\mathbb{E}\left[Y_{i,t=0}(0)|D_i = 0\right]\right), \quad (11) \end{aligned}$$

so that the percentage change in the means between $t = 0$ and $t = 1$ is the same for the treated and comparison groups. However, equation (11) is not equivalent to the parallel trends assumption in logs,

$$\mathbb{E}\left[log(Y_{i,t=1}(0))|D_i = 1\right] - \mathbb{E}\left[log(Y_{i,t=0}(0))|D_i = 1\right] \\ = \mathbb{E}\left[log(Y_{i,t=1}(0))|D_i = 0\right] - \mathbb{E}\left[log(Y_{i,t=0}(0))|D_i = 0\right], \quad (12)$$

which reverses the order of the logs and expectations. By Jensen's inequality, the order of the logs and expectations matters. Thus, while parallel trends in levels and equation (11) are incompatible when the $\mathbb{E}\left[Y_{i,t=s}(0) \mid D_i = d\right]$ are distinct, parallel trends in levels need not be incompatible with parallel trends in logs (or other transformations) if baseline levels differ.

Remark 13 (Different classes of transformations)**.** Following Athey and Imbens (2006), we define parallel trends to be invariant to transformations if it holds for all strictly monotonic (measurable) functions. It is straightforward to show, however, that if (2) holds, then parallel trends holds for all (measurable) g. Hence, parallel trends for all strictly monotonic functions is equivalent to parallel trends for *all* measurable functions. We note that this equivalence does not hold for other assumptions – e.g., the CiC model of Athey and Imbens (2006) is invariant to *strictly* monotonic transformations but not to all measurable transformations. We note further that once one requires parallel trends to hold for a "sufficiently rich" set of transformations, this will imply that it holds for all transformations. The reason for this is that if parallel trends holds for transformations g_1 and g_2, then it also holds for any affine combination of g_1 and g_2, and so invariance over a sufficiently rich set of transformations implies invariance over all transformations. For instance, if $Y_{i,t=s}(0)|D = d$ has a moment-generating function (MGF) for all (d, s), then it is sufficient to consider the set of exponential transformations $g_\lambda(y) = exp(\lambda y)$ for $\lambda \in \mathbb{R}$.[16] In Appendix B, we show that the distribution of untreated potential outcomes may only be partially identified if g is restricted to smaller classes of functions.

Remark 14 (Use of pre-treatment periods)**.** In settings where multiple pre-treatment periods are available, researchers may be inclined to use pre-treatment data to inform the choice of functional form. We note, however, that there may be many possible transformations g that satisfy parallel trends in the pre-treatment period. Indeed, we show in Appendix B that this will necessarily be the case when the support of $Y(0)$ is sufficiently rich. Thus, even if

[16]Specifically, parallel trends for this class of functions implies a "parallel trends"-type assumption for MGFs. Inverting both sides of the equation via inverse Laplace transforms then yields parallel trends of CDFs.

one is willing to impose that the "correct" functional form must satisfy parallel trends in the pre-treatment period, this will generally not be enough to point-identify the ATT without further assumptions.

4 Invariance of other estimators

The results in the previous section show that for parallel trends to be invariant to transformations, we require an assumption that pins down the entire distribution of counterfactual potential outcomes. A natural question is whether we might be able to construct a different estimator that allows for consistent estimation of the ATT for all monotonic transformations under weaker assumptions that do not pin down the full counterfactual distribution. The following result shows that the answer is no.

Proposition 4.1. *For any measurable function g, define*

$$\tau_{ATT}(g) = \mathbb{E}\left[g(Y_{i,t=1}(1)) - g(Y_{i,t=1}(0)) \mid D_i = 1\right].$$

Let $\mathcal{G}$ be the set of strictly monotonic, measurable functions g for which $\tau_{ATT}(g)$ is finite, and assume the identity map is in $\mathcal{G}$. Then $\tau_{ATT}(g)$ is identified for all $g \in \mathcal{G}$ if and only if $F^{Y(0)}_{D=1,t=1}(\cdot)$ is identified.

Proof. Suppose first $F^{Y(0)}_{D=1,t=1}(\cdot)$ is identified. Then $\mathbb{E}\left[g(Y_{i,t=1}(0)) \mid D_i = 1\right]$ is identified, since it is equal to $\int g(y) dF^{Y(0)}_{D=1,t=1}$. Further, $\mathbb{E}\left[g(Y_{i,t=1}(1)) \mid D_i = 1\right] = \mathbb{E}\left[g(Y_{i,t=1}) \mid D_i = 1\right]$, and thus is also identified. Hence $\tau_{ATT}(g) = \mathbb{E}\left[g(Y_{i,t=1}(1)) \mid D_i = 1\right] - \mathbb{E}\left[g(Y_{i,t=1}(0)) \mid D_i = 1\right]$ is identified.

Conversely, suppose $\tau_{ATT}(g)$ is identified for all $g \in \mathcal{G}$. By assumption, the identity map $g_1(y) = y$ is contained in $\mathcal{G}$. It follows that for any $\tilde{y} \in \mathbb{R}$, $g_2(y) = y - 1[y \leqslant \tilde{y}]$ is also contained in $\mathcal{G}$, since it is the sum of g_1 and a bounded, measurable function. Now,

$$\tau_{ATT}(g_1) - \tau_{ATT}(g_2) = \mathbb{E}\left[1[Y_{i,t=1}(1) \leqslant \tilde{y}] \mid D_i = 1\right] - \underbrace{\mathbb{E}\left[1[Y_{i,t=1}(0) \leqslant \tilde{y} \mid D_i = 1\right]}_{=F^{Y(0)}_{D=1,t=1}(\tilde{y})}$$

and hence

$$F^{Y(0)}_{D=1,t=1}(\tilde{y}) = \tau_{ATT}(g_2) - \tau_{ATT}(g_1) + \mathbb{E}\left[1[Y_{i,t=1}(1) \leqslant \tilde{y}] \mid D_i = 1\right].$$

However, the first two terms on the right-hand side of the previous display are identified by assumption, and the final term is equal to $\mathbb{E}\left[1[Y_{i,t=1} \leqslant \tilde{y}] \mid D_i = 1\right]$ and thus is identified. The result follows.

$\square$

Proposition 4.1 states that identification of $\tau_{ATT}(g)$ for all transformations g is equivalent to identification of the full distribution of untreated potential outcomes for the treated group. An immediate implication of this result is that there can exist a consistent or unbiased estimator of $\tau_{ATT}(g)$ for all g only if one imposes assumptions that identify the full distribution of untreated potential outcomes. Thus, the consistency or unbiasedness of an estimator for the ATT will necessarily be sensitive to functional form *unless* one imposes assumptions that identify the full distribution of counterfactual potential outcomes.

Remark 15 (Other classes of transformations)**.** Analogous to Remark 13, identification of $\tau_{ATT}(g)$ for a sufficiently rich set of g functions is equivalent to identification of $\tau_{ATT}(g)$ for all measurable functions. For instance, if $Y_{i,t=1}(d)|D = 1, t = 1$ has a moment generating function for $d = 0, 1$, then it suffices to consider the smaller set of transformations of the form $g_\lambda(y) = exp(\lambda y)$ for $\lambda \in \mathbb{R}$. Likewise, the result holds if one considers the larger set of all measurable transformations.

5 Discussion

Our theoretical results have important implications for practitioners who wish to obtain point-identification of the ATT using difference-in-differences or related research designs.

If treatment is (as-if) randomly assigned, then our results imply that the parallel trends assumption will hold regardless of the chosen functional form. Thus, sensitivity to functional form will not be an issue for DiD designs in settings with (as-if) random assignment. We note, however, that other estimators will also be valid under randomization of treatment and may be preferred over DiD on the basis of efficiency (McKenzie, 2012).

If treatment is not (as-if) randomly assigned, however, then our results imply that difference-in-differences will be sensitive to functional form unless the distributions of potential outcomes satisfy "parallel trends of CDFs" (Proposition 3.1). As discussed in Section 3 above, this assumption will generally be stronger than the usual parallel trends assumption in settings where treatment is not as-if randomly assigned. Indeed, Proposition 3.2 shows that this condition can hold essentially only if the population can be partitioned into a sub-group with non-stationary potential outcomes that is effectively randomized between treatment and control, and sub-groups which are non-randomized into treatment and control but have stable distributions of potential outcomes over time.

If these conditions for the invariance of parallel trends are implausible, then a researcher using a DiD design in non-experimental contexts should justify the specific functional form that is chosen. One way to do this is to argue for a specific model for the untreated potential outcomes. If context-specific knowledge motivates the model $\mathbb{E}\left[Y_{i,t}(0)|i,t=s\right] = \alpha_i + \lambda_s$, then parallel trends in levels is appropriate. Alternatively, if context-specific knowledge motivates the model $\mathbb{E}\left[\log(Y_{i,t}(0))|i,t=s\right] = \alpha_i + \lambda_s$, then parallel trends in logs may be reasonable. It is important to note that without strong assumptions, the choice of transformation on the left-hand side will matter, and so the researcher taking this approach must be careful to justify the chosen functional form at the exclusion of others.

Alternatively, the researcher might instead consider a different approach for modeling the entire distribution of untreated potential outcomes. For example, Athey and Imbens (2006) propose a model for inferring the distribution of counterfactual outcomes which does not depend on the functional form chosen for the outcome.[17] The Athey and Imbens (2006) approach is based on a model of heterogeneity that is non-nested with the assumptions needed for parallel trends to be invariant to transformations, and thus either might be preferred over the other depending on context.

In many settings, however, it will not be obvious which model for inferring the counterfactual distributions is correct. Unfortunately, Proposition 4.1 implies that it is not possible to point-identify the ATT without either taking a stand on functional form *or* taking a stand on how to infer the entire distribution of counterfactual outcomes.

If none of the assumptions needed to point-identify the ATT is plausible, researchers might instead impose weaker sets of assumptions that only partially identify the ATT. Manski and Pepper (2018) and Rambachan and Roth (2020b) consider partial identification under restrictions on the extent to which parallel trends can be violated for a particular functional form. These approaches relax functional form restrictions in that they only require parallel trends to hold approximately, rather than exactly, for a particular functional form, but they nevertheless require the researcher to specify a baseline functional form. An alternative approach would be to consider partial identification of the ATT under the assumption that parallel trends holds for some function g within a restricted class of transformations $\mathbf{G}$. We are not aware of any previous papers that consider such an approach, although we show in Appendix D that when $\mathbf{G}$ is a set of smooth functions, the restriction that parallel trends holds for $g \in \mathbf{G}$ implies restrictions of the form considered in Manski and Pepper (2018) and Rambachan and Roth (2020b). We think that further developing tools for partial identification under other classes of transformations is an interesting topic for future work.

[17] As mentioned in Remarks 9 and 10, Bonhomme and Sauder (2011) and Callaway and Li (2019) also propose models for inferring the counterfactual distribution of the outcome. However, their modelling assumptions rely on the chosen functional form for the potential outcomes $Y(0)$, so researchers using these models must again be careful about justifying choices of functional form.

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Almond, Douglas, Hilary W. Hoynes, and Diane Whitmore Schanzenbach, "Inside the War on Poverty: The Impact of Food Stamps on Birth Outcomes," *Review of Economics and Statistics*, May 2011, *93* (2), 387–403.

Andrews, Donald W. K. and Xiaoxia Shi, "Inference Based on Conditional Moment Inequalities," *Econometrica*, 2013, *81* (2), 609–666.

Angrist, Joshua D. and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

Angrist, Joshua D and Jörn-Steffen Pischke, "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, May 2010, *24* (2), 3–30.

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

_ and _ , "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *arXiv:1808.05293 [cs, econ, math, stat]*, August 2018.

Bonhomme, Stéphane and Ulrich Sauder, "Recovering Distributions in Difference-in-Differences models: A Comparison of Selective and Comprehensive Schooling," *Review of Economics and Statistics*, 2011, *93* (May), 479–494.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2020, (Forthcoming).

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

_ , _ , and Tatsushi Oka, "Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods," *Journal of Econometrics*, 2018, *206* (2), 395–413.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1405–1454. Publisher: Oxford Academic.

Cunha, Flavio and James J. Heckman, "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, October 2008, *43* (4), 738–782. Publisher: University of Wisconsin Press.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, September 2020, *110* (9), 2964–2996.

Delgado, Miguel and Juan Carlos Escanciano, "Conditional stochastic dominance testing," *Journal of Business & Economic Statistics*, 2013, *31* (1), 16–28.

Ding, Peng and Fan Li, "A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment," *Political Analysis*, 2019, *27* (4), 605–615.

Dube, Arindrajit, "Minimum Wages and the Distribution of Family Incomes," *American Economic Journal: Applied Economics*, October 2019, *11* (4), 268–304.

_ **and Attila S. Lindner**, "City Limits: What do Local-Area Minimum Wages Do?," *Journal of Economic Perspectives*, 2020, (Forthcoming).

Fang, Zheng, "Refinements of the Kiefer-Wolfowitz theorem and a test of concavity," *Eletronic Journal of Statistics*, 2019, *13*, 4596–4645.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, October 1997, *64* (4), 605–654. Publisher: Oxford Academic.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1 edition ed., New York: Cambridge University Press, April 2015.

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications," *Journal of Business & Economic Statistics*, 2020, *38* (3), 613–620.

Leamer, Edward E., "Let's Take the Con Out of Econometrics," *The American Economic Review*, 1983, *73* (1), 31–43. Publisher: American Economic Association.

_ , "Sensitivity Analyses Would Help," *The American Economic Review*, 1985, *75* (3), 308–313. Publisher: American Economic Association.

Lechner, Michael, "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 2011, *4*, 165–224.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, April 2015, *124*, 1–17.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

McKenzie, David, "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 2012, *99* (2), 210–221. Publisher: Elsevier.

Meyer, Bruce D., "Natural and Quasi-Experiments in Economics," *Journal of Business & Economic Statistics*, 1995, *13* (2), 151–161. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602 [econ, stat]*, August 2020. arXiv: 2008.00602.

_ and _ , "An Honest Approach to Parallel Trends," *Working Paper*, 2020.

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *Working paper*, 2020.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, November 2020, *219* (1), 101–122.

Stepner, Michael, "The Insurance Value of Redistributive Taxes and Transfers," 2019.

Sun, Liyan and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2020, (Forthcoming).

A Additional Proofs

A.1 Proof of Proposition 3.2

Proof. Observe that if (4) holds, then both sides of (2) reduce to $\theta(F_{t=1}(y) - F_{t=0}(y))$, and so (4) implies (2). To prove the converse, let $\mathcal{Y}$ denote the parameter space for $Y(0)$, and

$\mathcal{Y}_y = \{\tilde{y} \in \mathcal{Y} \mid \tilde{y} \leqslant y\}$. By assumption, we can write

$$F^{Y(0)}_{D=d,t=s}(y) = \int_{\mathcal{Y}_y} f_{D=d,t=s}\, d\lambda,$$

where λ is the dominating measure and $f_{D=d,t=s}$ is the density (the Radon-Nikodym derivative). It is immediate from the previous display that if (2) holds, then $f_{D=1,t=1} - f_{D=1,t=0} = f_{D=0,t=1} - f_{D=0,t=0}$, λ a.e. To prove that (2) implies (4), it thus suffices to establish the following claim:

Suppose the CDFs F_1 and F_2 are such that $F_j(y) = \int_{\mathcal{Y}_y} f_j d\lambda$. Then we can decompose $F_j(y)$ as

$$F_j(y) = \theta F_{min}(y) + (1-\theta)\tilde{F}_j(y), \tag{13}$$

where F_{min} and $\tilde{F}_1, \tilde{F}_2$ are CDFs, $\theta \in [0,1]$, and θ and $\tilde{F}_j$ depend on f_1 and f_2 only through $f_1 - f_2$.

To prove the claim, set $\theta = \int_{\mathcal{Y}} \min\{f_1, f_2\} d\lambda$. It is immediate that $\theta \in [0,1]$. Suppose first that $\theta \in (0,1)$. Define

$$f_{min} = \frac{\min\{f_1, f_2\}}{\int_{\mathcal{Y}} \min\{f_1, f_2\} d\lambda} = \frac{\min\{f_1, f_2\}}{\theta}$$

and

$$\tilde{f}_j(x) = \frac{f_j - \min\{f_1, f_2\}}{\int_{\mathcal{Y}} (f_j - \min\{f_1, f_2\}) d\lambda} = \frac{f_j - \min\{f_1, f_2\}}{1-\theta} \text{ for } j = 1, 2.$$

By construction, f_{min} and the $\tilde{f}_j$ integrate to 1 and are non-negative, so that $F_{min}(y) = \int_{\mathcal{Y}_y} f_{min} d\lambda$ and $\tilde{F}_j(y) = \int_{\mathcal{Y}_y} \tilde{f}_j d\lambda$ are valid CDFs. Moreover, $f_j = \theta f_{min} + (1-\theta)\tilde{f}_j$ by construction, so that (13) holds. Finally, observe that $\min\{f_1, f_2\} = f_1 - (f_1 - f_2)_+$, where $(a)_+$ denotes the positive part of a. It follows that $\theta = \int_{\mathcal{Y}} (f_1 - (f_1 - f_2)_+) d\lambda = 1 - \int_{\mathcal{Y}} (f_1 - f_2)_+ d\lambda$, which depends only on $f_1 - f_2$. (In fact, note that $\int_{\mathcal{Y}} (f_1 - f_2)_+ d\lambda = \frac{1}{2} \int_{\mathcal{Y}} |(f_1 - f_2)| d\lambda$, and thus θ is one minus the total variation distance between f_1 and f_2.) Likewise, $\tilde{f}_1 = (f_1 - f_2)_+/(1-\theta)$ and $\tilde{f}_2 = (f_2 - f_1)_+/(1-\theta)$, and so depend only on $f_1 - f_2$. This completes the proof for the case where $\theta \in (0,1)$. If $\theta = 1$, then $F_1(y) = F_2(y)$ and so the claim holds trivially with $F_{min}(y) = F_1(y) = F_2(y)$ and $\tilde{F}_j(y)$ arbitrary. If $\theta = 0$, then $\min\{f_1, f_2\} = 0$ λ a.e, and so $f_1 = (f_1 - f_2)_+$ λ a.e, and $f_2 = (f_2 - f_1)_+$ λ a.e. Thus, the claim holds trivially with $\tilde{f}_1 = f_j$, $\tilde{F}_j(y) = \int_{\mathcal{Y}_y} \tilde{f}_j d\lambda$, and $F_{min}(y)$ arbitrary.

□

A.2 Proof of impossibility result for normal CDFs

We claimed in Remark 6 that it is impossible for (6) to hold for all y if $(\mu_1, ...\mu_4)$ are distinct. Differentiating both sides of the equation, we see that if (6) holds for all y then

$$\phi_{\mu_1,\sigma_1}(y) = \phi_{\mu_2,\sigma_2}(y) + \phi_{\mu_3,\sigma_3}(y) - \phi_{\mu_4,\sigma_4}(y)$$

for all y. The following two lemmas show that this is impossible when the μ_j are distinct.

Lemma A.1. *The PDF of the $\mathcal{N}(\mu, \sigma^2)$ distribution is $\phi_{\mu,\sigma}(y) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(y-\mu)^2)$. We claim that:*
(i) If $0 < \sigma_1 < \sigma_2$, then $\lim_{y\to\infty} \dfrac{\phi_{\mu_1,\sigma_1}(y)}{\phi_{\mu_2,\sigma_2}(y)} = 0$ for any μ_1, μ_2.
(ii) If $\mu_1 < \mu_2$, then $\lim_{y\to\infty} \dfrac{\phi_{\mu_1,\sigma}(y)}{\phi_{\mu_2,\sigma}(y)} = 0$ for any $\sigma > 0$.

Proof. To prove (i), note that the ratio of interest can be written as $\frac{\sigma_2}{\sigma_1} exp(-\frac{1}{2\sigma_1^2}(y-\mu_1)^2 + \frac{1}{2\sigma_2^2}(y-\mu_2)^2)$. Note, however, that

$$\frac{\frac{1}{2\sigma_1^2}(y-\mu_1)^2}{\frac{1}{2\sigma_2^2}(y-\mu_2)^2} = \frac{\sigma_2^2}{\sigma_1^2}\left(\frac{y-\mu_1}{y-\mu_2}\right)^2 \to \frac{\sigma_2^2}{\sigma_1^2} > 1.$$

It follows that $-\frac{1}{2\sigma_1^2}(y-\mu_1)^2 + \frac{1}{2\sigma_2^2}(y-\mu_2)^2 \to -\infty$, which gives the desired result. To prove (ii), note that the ratio of interest can be written as $exp(-\frac{1}{2\sigma^2}(y-\mu_1)^2 + \frac{1}{2\sigma^2}(y-\mu_2)^2)$. But $(y-\mu_2)^2 - (y-\mu_1)^2 \to -\infty$ since $\mu_1 < \mu_2$, which gives the desired result. □

Lemma A.2. *Consider $Y_j \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right)$ for $j = 1, ..., 4$. Suppose $\mu_1, ..., \mu_4$ are distinct. Let $\sigma_1, ..., \sigma_4 > 0$ be arbitrary (not necessarily distinct). Then there exists y such that $\phi_{\mu_1,\sigma_1}(y) - \phi_{\mu_2,\sigma_2}(y) \neq \phi_{\mu_3,\sigma_3}(y) - \phi_{\mu_4,\sigma_4}(y)$.*

Proof. If not, then

$$\phi_{\mu_1,\sigma_1}(y) - \phi_{\mu_2,\sigma_2}(y) = \phi_{\mu_3,\sigma_3}(y) - \phi_{\mu_4,\sigma_4}(y)$$

for every y. If there is a unique maximum to $\{\sigma_1, .., \sigma_4\}$, let j^* be the index with maximal variance. Otherwise, let j^* be the index with the smallest μ_j among the j with maximal variance. j^* is unique by the assumption that the μ_j are distinct. From the previous lemma, $\lim_{y\to\infty} \dfrac{\phi_{\mu_j,\sigma_j}(y)}{\phi_{\mu_{j*},\sigma_{j*}}(y)}$ is 0 for $j \neq j^*$ and 1 for $j = j^*$. Dividing both sides of the previous display by by $\phi_{\mu_{j*},\sigma_{j*}}$ and taking limits as $y \to \infty$, it follows that one of the sides of the equality converges to 0 and the other converges to either positive or negative 1, which is a contradiction. □

B Extensions to Restricted Classes of Transformations and Learning from Pre-treatment Data

We now consider two extensions to the model considered in the main text. First, we consider restricted classes of functions $\mathbf{G}$ that may be smaller than the full set of strictly monotonic functions. Second, we consider the extent to which one can "learn" the right functional form from pre-treatment data. For ease of exposition, we consider a simplification of the main model with finite support.

B.1 Finite Support Set-Up

Suppose that $Y(0)$ has finite support, $\mathcal{Y} = \{y_1, ..., y_K\}$. Then the distribution $Y_{i,t=s}(0)|D = d$ is characterized by the k-dimensional vector

$$p_{ds} = \left(\mathbb{P}\left(Y_{i,t=s}(0) = y_1|D = d\right), ..., \mathbb{P}\left(Y_{i,t=s}(0) = y_K|D = d\right)\right)'.$$

If we define the vector $g = (g(y_1), ..., g(y_k))'$ for some function $g(\cdot)$, then our notation implies that

$$\mathbb{E}\left[g(Y_{i,t=s}(0))|D = d\right] = p'_{ds}g.$$

Hence, the usual parallel trends assumption between period $t = 0$ and $t = 1$ holds for a particular transformation $g(\cdot)$ if and only if

$$((p_{11} - p_{10}) - (p_{01} - p_{00}))'g = 0. \tag{14}$$

B.2 Restricted Classes of Transformations

Now, suppose we want (14) to hold for all $g \in \mathbf{G}$. Then, from equation (14), we see that $\tilde{p} := ((p_{11} - p_{10}) - (p_{01} - p_{00}))$ must lie in the null space of the linear subspace generated by $\mathbf{G}$. If $\mathbf{G}$ is sufficiently rich that it spans $\mathbb{R}^K$, then we must have that $\tilde{p} = 0$, i.e. parallel trends of PMFs (which is equivalent to parallel trends of CDFs). If $\mathbf{G}$ is less rich, with say span of dimension $m < K$, then (14) implies only the weaker condition that $\tilde{p}$ lies in the $K - m$ dimensional nullspace of $\mathbf{G}$. Hence, the difference-in-difference of PMFs is partially identified for smaller classes of transformations $\mathbf{G}$.

B.3 Learning g from pre-treatment periods

Likewise, if we want pre-treatment parallel trends to hold for periods $t = -T, ..., 0$, then we require that

$$\tilde{P}g = 0,$$

where $\tilde{P}$ is the $T \times K$ matrix with jth row equal to $(p_{1,-j+1} - p_{1,-j}) - (p_{0,-j+1} - p_{0,-j})$.

We might hope that we can identify the "right" transformation g by requiring it to satisfy the pre-treatment version of parallel trends, $\tilde{P}g = 0$. Unfortunately, the following result shows that this often will not be enough to pin down the correct functional form. In particular, if the number of support points for $Y(0)$ is larger than the number of pre-treatment periods plus two, then the transformation satisfying pre-treatment parallel trends (if it exists) will not be unique. Moreover, there will be distributions for $Y_{i,t=1}(0)$ for the treated group such that at least one of the transformations that satisfies parallel trends in the pre-treatment period fails in the post-treatment period. Thus, if the support of $Y(0)$ is sufficiently rich, then it is not enough to select a functional form that satisfies parallel trends in the pre-treatment period; one must either impose additional functional form assumptions or restrictions on the distribution.

Proposition B.1. *Suppose that $K > T + 2$. Then one of the following holds: 1) There is no strictly monotonic g such that $\tilde{P}g = 0$. 2) There exist two strictly monotonic vectors, g_1 and g_2, and a valid PMF p_{11}, such that such that both g_1 and g_2 satisfy pre-treatment parallel trends, $\tilde{P}g_j = 0$ for $j = 1, 2$, but parallel trends fails for at least one g_j, i.e. $((p_{11} - p_{10}) - (p_{01} - p_{00}))'g_j \neq 0$ for at least one j.*

Proof. Suppose there exists a strictly monotonic g_1 such that $\tilde{P}g_1 = 0$. We will then show that there exists a strictly monotonic g_2 and distribution p_{11} such that 2) holds.

First, we construct a strictly monotonic g_2 satisfying $\tilde{P}g_2 = 0$. Recall that $\tilde{P}$ is a $T \times K$ matrix. Since $T < K - 2$, the null-space of $\tilde{P}$ has dimension at least 3. Since p_{dt} is a PMF, it must sum to 1, i.e. $p'_{dt}\iota = 1$, where ι is the vector of 1s. It follows that $\tilde{P}\iota = 0$. Hence, there exists a non-zero vector v orthogonal to g_1 and ι such that $\tilde{P}v = 0$. Since g_1 is strictly monotonic, $g_2 := g_1 + c \cdot v$ is also strictly monotonic for $c > 0$ sufficiently small. Without loss of generality, assume that this holds for $c = 1$, so that $g_2 = g_1 + v$. The vector g_2 satisfies $\tilde{P}v = 0$ by construction.

We now construct a p_{11} such that parallel trends fails for at least one g_j. Observe that parallel trends must fail for at least one g_j if

$$((p_{11} - p_{10}) - (p_{01} - p_{00}))' \underbrace{(g_2 - g_1)}_{=v} \neq 0. \tag{15}$$

Next, for any $\epsilon \in (0,1)$, let

$$p_{11}^{\epsilon} = \epsilon v + \frac{1}{K}\iota.$$

Observe that $\iota' p_{11}^{\epsilon} = \frac{1}{K}\iota'\iota = 1$, since by construction v is orthogonal to ι. Hence, p_{11}^{ϵ} is a valid PMF if its minimal entry is non-negative. However, the minimal entry is given by $\epsilon \min_j v_j + \frac{1}{K}$, and hence there exists $\bar{\epsilon} > 0$ such that p_{11}^{ϵ} is a valid PMF for all $\epsilon \in (0, \bar{\epsilon})$. But for $0 < \epsilon_1 < \epsilon_2 < \bar{\epsilon}$, we have that $(p_{11}^{\epsilon_2} - p_{11}^{\epsilon_1}) = (\epsilon_2 - \epsilon_1)v$, and hence $(p_{11}^{\epsilon_2} - p_{11}^{\epsilon_1})'v \neq 0$. It follows that (15) holds for at least one of $p_{11}^{\epsilon_1}$ or $p_{11}^{\epsilon_2}$, which completes the proof. $\square$

C Values of θ for which Proposition 3.2 holds

We show in Proposition 3.2 that parallel trends of CDFs holds if and only if $F_{D=d,t=s}^{Y(0)}$ satisfies the decomposition

$$F_{D=d,t=s}^{Y(0)}(y) = \theta F_{t=s}^{Y(0)}(y) + (1-\theta) F_{D=d}^{Y(0)}(y) \text{ for all } y \in \mathbb{R} \text{ and } d, s \in \{0,1\}, \tag{16}$$

where $\theta \in [0,1]$ and $F_{t=s}^{Y(0)}(y)$ and $F_{D=d}^{Y(0)}(y)$ are CDFs of distributions that depend only on time and group, respectively. The following result characterizes the range of values of θ for which the decomposition above can be satisfied.

Proposition C.1. *Assume the measure theoretic condition in Proposition 3.2 holds. Let $F_{D=d,t=s}^{Y(0)} = \int f_{D=d,t=s} d\lambda$. Define θ_{lb} to the total variation distance between $Y(0) \mid D = 0, t = 1$ and $Y(0) \mid D = 0, t = 0$,*

$$\theta_{lb} = \int (f_{D=0,t=1} - f_{D=0,t=0})_+ d\lambda,$$

where $(x)_+ = max(x, 0)$. Likewise, define θ_{ub} to be one minus the total variation distance between $Y(0) \mid D = 0, t = 0$ and $Y(0) \mid D = 1, t = 0$,

$$\theta_{ub} = 1 - \int (f_{D=1,t=0} - f_{D=0,t=0})_+ d\lambda.$$

Suppose parallel trends of CDFs holds. Then for any $\theta \in [\theta_{lb}, \theta_{ub}]$ there exist distributions $F_{D=d}^{Y(0)}$ and $F_{t=s}^{Y(0)}$ such that (16) holds. Further, these bounds are sharp in the sense that there are no such distributions for any θ outside of this interval.

Proof. We first show that under parallel trends of CDFs, the decomposition (16) is satisfied for $\theta = \theta^{lb}$ and $\theta = \theta^{ub}$. In the proof to Proposition 3.2, we proved the following claim: suppose the CDFs F_1 and F_2 are such that $F_j(y) = \int_{\mathcal{Y}} f_j d\lambda$, $j = 1, 2$. Then we can decompose $F_j(y)$ as

$$F_j(y) = \tilde{\theta} F_{min}(y) + (1 - \tilde{\theta}) \tilde{F}_j(y),$$

where F_{min} and $\tilde{F}_1, \tilde{F}_2$ are CDFs, $\tilde{\theta} = 1 - \int (f_1 - f_2)_+ d\lambda \in [0, 1]$, and $\tilde{\theta}$ and $\tilde{F}_j$ depend on f_1 and f_2 only through $f_1 - f_2$.

Applying this claim on both sides of (2) gives the decomposition (16) with $\theta = \int (f_{D=0,t=1} - f_{D=0,t=0})_+ d\lambda = \theta_{lb}$.[18] However, rearranging the terms in parallel trends of CDFs, we have that

$$F^{Y(0)}_{D=1,t=1}(y) - F^{Y(0)}_{D=0,t=1}(y) = F^{Y(0)}_{D=1,t=0}(y) - F^{Y(0)}_{D=0,t=0}(y), \text{ for all } y \in \mathbb{R}.$$

Applying the claim to both sides then yields the desired decomposition with $\theta = \theta_{ub}$.

Next, we show that if (16) holds for θ_{lb} and θ_{lb}, then it also holds for any $\theta_r = r\theta_{lb} + (1 - r)\theta_{ub}$ for $r \in (0, 1)$. In particular, note that the decomposition (16) holds if and only if it holds at almost every y for the densities,

$$f_{D=d,t=s} = \theta f_{t=s} + (1 - \theta) f_{D=d}. \tag{17}$$

Thus, we have shown so far that (for almost every y)

$$f_{D=d,t=s} = \theta_{lb} f^{lb}_{t=s} + (1 - \theta_{lb}) f^{lb}_{D=d}$$

and

$$f_{D=d,t=s} = \theta_{ub} f^{ub}_{t=s} + (1 - \theta_{ub}) f^{ub}_{D=d}.$$

However, combining the equations in the previous two displays with weights r and $1 - r$ respectively yields that

$$f_{D=d,t=s} = \theta_r f^r_{t=s} + (1 - \theta_r) f^r_{D=d},$$

where

$$f^r_{t=s} = \frac{r\theta_{lb}}{\theta_r} f^{lb}_{t=s} + \frac{(1 - r)\theta_{ub}}{\theta_r} f^{ub}_{t=s},$$

and $f^r_{D=d}$ is defined analogously.

Finally, we show the bounds are tight. Note that equation (17) implies that

$$f_{D=0,t=1} - f_{D=0,t=0} = \theta(f_{t=1} - f_{t=0}).$$

Taking the positive part of both sides and integrating, we have that

$$\theta_{lb} = \int (f_{D=0,t=1} - f_{D=0,t=0})_+ d\lambda = \theta \int (f_{t=1} - f_{t=0})_+ d\lambda \leqslant \theta,$$

where the inequality follows from the fact that $0 \leqslant \int (f_{t=1} - f_{t=0})_+ d\lambda \leqslant \int f_{t=1} d\lambda = 1$. We have thus shown that the lower bound is tight. The proof for the upper bound is analogous. ☐

[18] By parallel trends of CDFs, this is also equal to $\int (f_{D=1,t=1} - f_{D=1,t=0})_+ d\lambda$.

An immediate corollary of this result is that there exists some distribution for $Y(0) \mid D = 1, t = 1$ such that parallel trends of CDFs holds (or equivalently, parallel trends is invariant to transformations) if and only if the identified distributions of $Y(0) \mid D = d, t = s$ are "close enough" in either the time or group dimension, where "close enough" is determined by the total variation distance.

Corollary C.1. *Assume the measure theoretic condition in Proposition 3.2 holds. Then there exists a distribution for $Y(0) \mid D = 1, t = 1$ such that parallel trends of CDFs is satisfied if and only if the sum of i) the total variation distance between $Y(0) \mid D = 0, t = 0$ and $Y(0) \mid D = 1, t = 0$, and ii) the total variation distance between $Y(0) \mid D = 0, t = 0$ and $Y(0) \mid D = 0, t = 1$, is less than or equal to 1.*

D Partial Identification of the ATT

If the assumptions needed to point-identify the ATT are unappealing, one can instead rely on weaker assumptions that imply partial identification of the ATT. There are two approaches to partial identification: first, one might place bounds on the extent to which parallel trends might fail for a particular functional form. Second, one might impose that parallel trends holds exactly for *some* transformation g in a set of possible classes of transformations. We now briefly describe each approach.

We begin with the approach of bounding the extent to which parallel trends can fail, as considered in Manski and Pepper (2018); Rambachan and Roth (2020b). Let δ_1 denote the difference in trends in untreated potential outcomes:

$$\delta_1 := \left(\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 1\right]\right) - \left(\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 0\right]\right).$$

Manski and Pepper (2018) and Rambachan and Roth (2020b) consider identification and inference for τ_{ATT} under assumptions that bound the possible magnitude of δ_1. Rambachan and Roth (2020b) also consider assumptions that restrict the possible magnitude of δ_1 in terms of other quantities that are identified from the data, e.g. pre-treatment differences in trends. These approaches only require that parallel trends in levels not fail too badly – and facilitate sensitivity analysis with respect to the magnitude of the potential failures – and thus may be appealing in contexts where the researcher is unsure whether a particular functional form (say, levels) is exactly right.

We next consider identification of the ATT in levels under the assumption that parallel trends holds exactly for some $g \in \mathbf{G}$. We are not aware of any previous papers considering such restrictions, but we will show that when $\mathbf{G}$ is a set of smooth functions, the restriction that $g \in \mathbf{G}$ implies restrictions of the form considered in Manski and Pepper (2018);

Rambachan and Roth (2020b).

In particular, note that the usual parallel trends assumption corresponds with imposing that parallel trends holds for any affine function g. A natural relaxation of this restriction would therefore be to impose that parallel trends holds for some "smooth" class of functions $\mathbf{G}$. In particular, consider the set of transformations with second derivative bounded by M, $\mathbf{G}_{Holder}(M) = \{g : g'(0) = 1, g''(x) \leqslant M \, \forall x\}$.[19] Such smoothness restrictions are common in non-parametric regression, e.g., Armstrong and Kolesar (2018); Kolesar and Rothe (2018); Frandsen (2016); Noack and Rothe (2020). The following lemma shows that if $Y(0)$ has bounded support, then the assumption that parallel trends holds for some $g \in \mathbf{G}_{Holder}(M)$ implies restrictions on the possible magnitude of δ_1. Restricting g to $\mathbf{G}_{Holder}(M)$ thus implies restrictions of the form considered in Manski and Pepper (2018) and Rambachan and Roth (2020b).

Lemma D.1. *Suppose the support of $Y(0)$ is contained within $[-C, C]$ for some finite C. If parallel trends holds for some $g \in \mathbf{G}_{Holder}(M)$, then $|\delta_1| \leqslant 2MC^2$.*

Proof. For any $y \in [-C, C]$, we can write $g(y) = g(0) + y + \frac{1}{2}g''(\tilde{y})y^2$ for some $\tilde{y}$ between 0 and y. Thus,

$$\mathbb{E}\left[g(Y_{i,t=s}(0)) \mid D_i = d\right] = \mathbb{E}\left[g(0) + Y_{i,t=s}(0) + g''(\tilde{Y}_{i,t=s})Y_{i,t=s}(0)^2 \mid D_i = d\right],$$

where $\tilde{Y}_{i,t=s}$ is now a random variable depending on $Y_{i,t=s}(0)$. It follows that

$$\begin{aligned}
&(\mathbb{E}\left[g(Y_{i,t=1}(0)) \mid D_i = 1\right] - \mathbb{E}\left[g(Y_{i,t=0}(0)) \mid D_i = 1\right]) - \\
&\quad (\mathbb{E}\left[g(Y_{i,t=1}(0)) \mid D_i = 0\right] - \mathbb{E}\left[g(Y_{i,t=0}(0)) \mid D_i = 0\right]) \\
&= (\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 1\right]) - \\
&\quad (\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 0\right]) + \\
&\quad \frac{1}{2}\left(\mathbb{E}\left[g''(\tilde{Y}_{i,t=1})Y_{i,t=1}(0)^2 \mid D_i = 1\right] - \mathbb{E}\left[g''(\tilde{Y}_{i,t=0})Y_{i,t=0}(0)^2 \mid D_i = 1\right]\right) - \\
&\quad \frac{1}{2}\left(\mathbb{E}\left[g''(\tilde{Y}_{i,t=1})Y_{i,t=1}(0)^2 \mid D_i = 0\right] - \mathbb{E}\left[g''(\tilde{Y}_{i,t=0})Y_{i,t=0}(0)^2 \mid D_i = 0\right]\right)
\end{aligned}$$

If parallel trends holds for some $g \in \mathbf{G}_{Holder}(M)$, then it follows that

$$\begin{aligned}
&|(\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 1\right]) - \\
&\qquad\qquad (\mathbb{E}\left[Y_{i,t=1}(0) \mid D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}(0) \mid D_i = 0\right])|
\end{aligned}$$

[19] We impose the additional normalization that $g'(0) = 1$. The reason for this is that if parallel trends holds for a transformation g, then it also holds for $\tilde{g}(y) = \epsilon g(y)$ for any ϵ. Thus, if we did not impose the normalization that $g'(0) = 1$, then imposing that parallel trends holds for all $g \in \mathbf{G}_{Holder}(M)$ would require that parallel trends hold for all g with bounded second derivative regardless of the value of M.

$$
\begin{aligned}
\leqslant & \frac{1}{2}\left|\left(\mathbb{E}\left[g''(\tilde{Y}_{i,t=s})Y_{i,t=s}(0)^2 \mid D_i = 1\right] - \mathbb{E}\left[g''(\tilde{Y}_{i,t=s})Y_{i,t=s}(0)^2 \mid D_i = 1\right]\right) - \right. \\
& \qquad\qquad \left.\left(\mathbb{E}\left[g''(\tilde{Y}_{i,t=s})Y_{i,t=s}(0)^2 \mid D_i = 0\right] - \mathbb{E}\left[g''(\tilde{Y}_{i,t=s})Y_{i,t=s}(0)^2 \mid D_i = 0\right]\right)\right| \\
\leqslant & \frac{1}{2}\sum_{d,s}\left|\mathbb{E}\left[g''(\tilde{Y}_{i,t=s})Y_{i,t=s}(0)^2 \mid D_i = d\right]\right| \\
\leqslant & 2MC^2.
\end{aligned}
$$

□

In light of Lemma D.1, one can use the inference tools put forward by Rambachan and Roth (2020b) when the goal is to partially identify the ATT in levels and one only imposes that parallel trends holds for some $g \in \mathbf{G}_{Holder}(M)$. Extending these results to other classes of functions is an interesting topic for future research.

E Design-Based Results

In the main text to the paper, we considered identification from a super-population perspective. We now show that similar results hold from a design-based perspective in which the randomness in the data arises from the stochastic treatment assignment, and the units in the population and their potential outcomes are treated as fixed.

E.1 Model

We consider a two-period difference-in-differences model from a design-based perspective, as in Athey and Imbens (2018); Rambachan and Roth (2020a). There is a finite population of N units. We observe data for 2 periods $t = 0, 1$. All units are untreated in $t = 0$, and some units receive a treatment of interest in $t = 1$. We denote by $Y_{i,t}(1), Y_{i,t}(0)$ the potential outcomes for unit i in period t under treatment and control, respectively, and we observe the outcome $Y_{i,t} = D_i Y_{i,t}(1) + (1 - D_i)Y_{i,t}(0)$, where D_i is an indicator for whether unit i is treated. We assume that there are no anticipatory effects of treatment, so that $Y_{i,t=0}(1) = Y_{i,t=0}(0)$ for all i. Following Neyman (1923) and Fisher (1935) for randomized experiments and Athey and Imbens (2018); Rambachan and Roth (2020a) for DiD designs, we treat as fixed (or condition on) the potential outcomes and the number of treated and untreated units (N_0 and N_1). The only source of uncertainty in our model comes from the vector of treatment assignments $\mathbf{D}$, which is probabilistic.

For notation, we will define $\mathbf{D} = (D_1, ..., D_n)'$ to be the (random) vector of treatment assignments. We condition on the number of treated units N_1, so the support of $\mathbf{D}$ is

$\{d \in \{0,1\}^n \mid \sum_i d_i = N_1\}$. We denote by $\mathbf{Y}_t(d) = (Y_{1t}(d), ..., Y_{Nt}(d))'$ the vector of potential outcomes for $t, d \in \{0,1\}$, $\mathbf{Y}_t = D\mathbf{Y}_t(1) + (1-D)\mathbf{Y}_t(0)$ the vector of realized outcomes in period t, and $\mathbf{Y} = (\mathbf{Y}_0', \mathbf{Y}_1')'$ and $\mathbf{Y}(\bullet) = (\mathbf{Y}_1(1)', \mathbf{Y}_0(1)', \mathbf{Y}_1(0)', \mathbf{Y}_0(0)')'$ the stacked vectors of realized and potential outcomes. All expectations and probability statements are taken over the distribution of D conditional on $(\mathbf{Y}(\bullet), N_0, N_1)$, although we will suppress this conditioning in our notation unless needed for clarity. We denote by P_D the probability distribution of D conditional on $(\mathbf{Y}(\bullet), N_0, N_1)$, and refer to this as the assignment mechanism. We will define $\pi_i := \mathbb{P}_{P_D}(D_i = 1)$ to be the marginal probability that i is treated under P_D.

E.2 Analysis of Difference-in-Differences

We now consider the properties of the canonical difference-in-differences estimator,

$$\hat{\tau}^{DiD}(\mathbf{D}, \mathbf{Y}) = \frac{1}{N_1} \sum_i D_i (Y_{i,t=1} - Y_{i,t=0}) - \frac{1}{N_0} \sum_i (1 - D_i)(Y_{i,t=1} - Y_{i,t=0}). \tag{18}$$

We write the estimator explicitly as a function of the observed data $(\mathbf{D}, \mathbf{Y})$, although we suppress this notation when it is not needed for clarity.

E.2.1 Parallel Trends and Unbiasedness

We first consider the expectation of $\hat{\tau}^{DiD}$. Let $\tau_i(\mathbf{Y}(\bullet)) = Y_{i,t=1}(1) - Y_{i,t=1}(0)$ be i's treatment effect in $t = 1$. We write τ_i explicitly as a function of the potential outcomes, since we will be interested in causal effects under different transformations of the potential outcomes. Define the average treatment effect on the treated (ATT) by

$$\tau_{ATT}(\mathbf{Y}(\bullet)) := \mathbb{E}_{P_D}\left[\sum_i \frac{1}{N_1} D_i \tau_i(\mathbf{Y}(\bullet))\right] = \frac{1}{N_1} \sum_i \pi_i \tau_i(\mathbf{Y}(\bullet)).$$

It is straightforward to show that $\hat{\tau}^{DiD}$ is an unbiased estimator of $\tau_{ATT}(\mathbf{Y}(\bullet))$ if and only if a parallel trends assumption holds, meaning that

$$\mathbb{E}_{P_D}\left[\frac{1}{N_1} \sum_i D_i (Y_{i,t=1}(0) - Y_{i,t=0}(0))\right] - \mathbb{E}_{P_D}\left[\frac{1}{N_0} \sum_i (1 - D_i)(Y_{i,t=1}(0) - Y_{i,t=0}(0))\right] = 0. \tag{19}$$

Lemma E.1. *The following are equivalent:*

(1) $\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}\right] = \tau_{ATT}(\mathbf{Y}(\bullet))$.

(2) Parallel trends (equation (19)) holds.

(3) $\sum_i \dot{\pi}_i (Y_{i,t=1}(0) - Y_{i,t=0}(0)) = 0$, *where* $\dot{\pi}_i = \pi_i - \frac{N_1}{N}$.

Proof. To obtain the equivalence of (1) and (2), observe that

$$\begin{aligned}
&\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}\right] \\
&= \mathbb{E}_{P_D}\left[\frac{1}{N_1}\sum_i D_i(Y_{i,t=1}(1) - Y_{i,t=0}(1))\right] - \mathbb{E}_{P_D}\left[\frac{1}{N_0}\sum_i (1-D_i)(Y_{i,t=1}(0) - Y_{i,t=0}(0))\right] \\
&= \mathbb{E}_{P_D}\left[\frac{1}{N_1}\sum_i D_i \tau_i\right] + \\
&\quad \mathbb{E}_{P_D}\left[\frac{1}{N_1}\sum_i D_i(Y_{i,t=1}(0) - Y_{i,t=0}(0))\right] - \mathbb{E}_{P_D}\left[\frac{1}{N_0}\sum_i (1-D_i)(Y_{i,t=1}(0) - Y_{i,t=0}(0))\right],
\end{aligned}$$

where the second line uses the fact that $Y_{i,t=0}(1) = Y_{i,t=0}(0)$ by assumption and adds and subtracts terms. The equivalence between (2) and (3) then follows from the fact that $\mathbb{E}_{P_d}[D_i] = \pi_i$ and $\pi_i/N_1 - (1-\pi_i)/N_0 = (N/(N_1 N_0))\dot{\pi}_i$. □

E.2.2 Parallel trends for all potential outcomes

We first consider when the parallel trends assumption depends only on the assignment mechanism P_D and not on the potential outcomes. The following result is an immediate corollary of Lemma E.1.

Corollary E.1. *Parallel trends holds for all* $\mathbf{Y}(\bullet)$ *if and only if* P_D *is such that* $\pi_i = \frac{N_1}{N}$ *for all* i.

Proof. From Lemma E.1, parallel trends holds if and only if $\sum_i \dot{\pi}_i (Y_{i,t=1}(0) - Y_{i,t=0}(0)) = 0$, where $\dot{\pi}_i = \pi_i - \frac{N_1}{N}$. Clearly, this holds if $\dot{\pi}_i \equiv 0$. Conversely, if $\dot{\pi}_i \neq 0$ for some i, then this is violated if we set the potential outcomes such that $Y_{i,t=1}(0) - Y_{i,t=0}(0)$ is proportional to π_i. □

Corollary E.1 makes clear that the parallel trends assumption places no restrictions on the potential outcomes if and only if treatment probabilities are equal for all units. This is guaranteed by design in random experiments, but often will be implausible in non-experimental settings. If different units have different probabilities of receiving treatment, then the parallel trends assumption will necessarily place some restrictions on the potential outcomes.

E.3 Invariance to transformations

We next consider the extent to which the parallel trends assumption imposes functional form restrictions on $\mathbf{Y}(\bullet)$. More concretely, we will again say that the parallel trends assumption is invariant to (monotonic) transformations if when parallel trends holds for potential outcomes $\mathbf{Y}(\bullet)$, it also holds if we replace $\mathbf{Y}(\bullet)$ with $g(\mathbf{Y}(\bullet))$ for any strictly monotonic transformation g, following Athey and Imbens (2006).

The following characterization shows that the parallel trends is invariant to transformations if and only if a "parallel trends"-type assumption holds on the entire cumulative distribution function (CDF) of untreated potential outcomes, and can be viewed as a design-based analog of Proposition 3.1.

Proposition E.1. *For any assignment mechanism P_D and vector of potential outcomes $\mathbf{Y}(\bullet)$, the following are equivalent:*

1. *Parallel trends is invariant to transformations, i.e.*

$$\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}(\mathbf{D}, g(\mathbf{Y}))\right] = \tau^{ATT}(g(\mathbf{Y}(\bullet))) \tag{20}$$

for all strictly monotonic functions g.

2. *For all y,*

$$F^{Y(0)}_{D=1,t=1}(y) - F^{Y(0)}_{D=1,t=0}(y) = F^{Y(0)}_{D=0,t=1}(y) - F^{Y(0)}_{D=0,t=0}(y), \tag{21}$$

where $F^{Y(0)}_{D=1,t=1}(y) = \mathbb{E}_{P_D}\left[\frac{1}{N_1}\sum_{\{i:D_i=1\}} 1[Y_{i,t=1}(0) \leqslant y]\right]$ is the (expected) CDF of the distribution of the untreated potential outcomes for treated units in period 1, and the other CDFs are defined analogously.

Proof. See Section E.5. □

Remark 16 (Random experiment)**.** If all units have the same treatment probabilities, $\pi_i \equiv \frac{N_1}{N}$, then equation (21) holds automatically, since $F^{Y(0)}_{D=1,t} = F^{Y(0)}_{D=0,t}$ for all t by virtue of random assignment. Indeed, we showed in Corollary E.1 that under random assignment, parallel trends does not depend at all on the potential outcomes, which is a stronger notion of robustness than invariance to transformations.

Remark 17 (Non-randomized settings)**.** Outside of randomized settings, however, the parallel trends of CDFs required in (21) will often be restrictive, for reasons similar to those discussed in Section 3 from the super-population view.

E.4 Extension to other estimators

The results in the previous section show that for parallel trends to be invariant to transformations, we require an assumption that pins down the entire distribution of counterfactual potential outcomes. A natural question is whether we might be able to construct a different estimator that achieves unbiased estimation of the ATT for all monotonic transformations under weaker assumptions that do not pin down the full counterfactual distribution. The following result shows that the answer is no, and is analogous to Proposition 4.1.

Proposition E.2. *Suppose that the assignment mechanism and* $\mathbf{Y}(\bullet)$ *are such that for all strictly monotonic functions* g*,*

$$\mathbb{E}_{P_D}\left[\hat{\tau}(\mathbf{D}, g(\mathbf{Y}))\right] = \tau_{ATT}(g(\mathbf{Y}(\bullet))) \tag{22}$$

for some estimator $\hat{\tau}$*. Then there exists a function* $\hat{F}^{Y(0)}_{D=1,t=1}(\cdot,\cdot;y)$ *such that*

$$\mathbb{E}_{P_D}\left[\hat{F}^{Y(0)}_{D=1,t=1}(\mathbf{D}, \mathbf{Y}; y)\right] = F^{Y(0)}_{D=1,t=1}(y)$$

for all y*.*

Proof. See Section E.5. □

Proposition E.2 shows that the existence of any estimator that is unbiased for the ATT under all monotonic transformations implies that the counterfactual distribution function $F^{Y(0)}_{D=1,t=1}$ is identified. It is thus not possible to obtain a robust estimator of the ATT without (implicitly) taking a stand on the entire distribution of counterfactual outcomes.

E.5 Additional Proofs for Design-Based Results

Proof of Proposition E.1

Proof. Let $\dot{p}_t(y) := \sum_{\{i: Y_{i,t}(0)=y\}} \dot{\pi}_i$ and let $\mathcal{Y} := \{\mathbf{Y}(\bullet)\}$ be the support of the potential outcomes. We will show that (20) and (21) are both equivalent to the following statement:

$$\dot{p}_1(y) = \dot{p}_0(y) \text{ for all } y \in \mathcal{Y}. \tag{23}$$

We first show that (20) is equivalent to (23). By Lemma E.1,

$$\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}(\mathbf{D}, g(\mathbf{Y}))\right] = \tau_{ATT}(g(\mathbf{Y}(\bullet))) \Leftrightarrow \frac{1}{N}\sum_i \dot{\pi}_i \left(g(Y_{i,t=1}(0)) - g(Y_{i,t=0}(0))\right) = 0.$$

Note that

$$\frac{1}{N}\sum_i \dot{\pi}_i g(Y_{i,t=1}(0)) = \frac{1}{N}\sum_{y\in\mathcal{Y}} \sum_{\{i: Y_{i,t=1}(0)=y\}} \dot{\pi}_i g(y) = \frac{1}{N}\sum_{y\in\mathcal{Y}} \dot{p}_1(y) g(y)$$
$$\frac{1}{N}\sum_i \dot{\pi}_i g(Y_{i,t=0}(0)) = \frac{1}{N}\sum_{y\in\mathcal{Y}} \sum_{\{i: Y_{i,t=0}(0)=y\}} \dot{\pi}_i g(y) = \frac{1}{N}\sum_{y\in\mathcal{Y}} \dot{p}_0(y) g(y).$$

Combining the previous three displays, we see that

$$\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}(\mathbf{D}, g(\mathbf{Y}))\right] = \tau_{ATT}(g(\mathbf{Y}(\bullet))) \Leftrightarrow \frac{1}{N}\sum_{y\in\mathcal{Y}}(\dot{p}_1(y) - \dot{p}_0(y))g(y) = 0. \tag{24}$$

It is immediate that (20) holds for all g if $\dot{p}_1(y) = \dot{p}_0(y)$. Conversely, suppose that $\dot{p}_1(y) \neq \dot{p}_0(y)$ for some $y \in \mathcal{Y}$. Let $\tilde{y} = \max\{y \in \mathcal{Y} : p_1(y) \neq p_0(y)\}$, which is finite since $\mathcal{Y}$ is finite. Let $g_1(y) = y$ for $y < \tilde{y}$ and $g_1(y) = \tilde{y} + 1$ for $y \geqslant \tilde{y}$. Using (24) and the fact that $\dot{p}_1(y) = \dot{p}_0(y)$ for all $y > \tilde{y}$ by construction, we have

$$\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}(\mathbf{D}, g_1(\mathbf{Y}))\right] = \tau_{ATT}(g_1(\mathbf{Y}(\bullet))) \Leftrightarrow \left(\frac{1}{N}\sum_{y\in\mathcal{Y}}(\dot{p}_1(y) - \dot{p}_0(y))y\right) + \dot{p}_1(\tilde{y}) - \dot{p}_0(\tilde{y}) = 0.$$

But applying (24) again using $g_2(y) = y$, we see that

$$\mathbb{E}_{P_D}\left[\hat{\tau}^{DiD}(\mathbf{D}, \mathbf{Y})\right] = \tau_{ATT}(\mathbf{Y}(\bullet)) \Leftrightarrow \frac{1}{N}\sum_{y\in\mathcal{Y}}(\dot{p}_1(y) - \dot{p}_0(y))y = 0.$$

It follows that (20) is violated for either g_1 or g_2.

Now, let

$$f^{Y(0)}_{D=1,t=1}(y) = \mathbb{E}_{P_D}\left[\frac{1}{N_1}\sum_{\{i:D_i=1\}} 1[Y_{i,t=1}(0) = y] \,|\, \mathbf{Y}(\bullet)\right],$$

and define $f_{D=0,t=1}$, $f_{D=1,t=0}$, and $f_{D=0,t=0}$ analogously. Note that

$$F^{Y(0)}_{D=d,t=t}(y) = \sum_{\tilde{y}\in\mathcal{Y},\tilde{y}\leqslant y} f^{Y(0)}_{D=d,t=t}(y),$$

from which it follows that (21) holds for all y if and only if

$$f^{Y(0)}_{D=1,t=1}(y) - f^{Y(0)}_{D=1,t=0}(y) = f^{Y(0)}_{D=0,t=1}(y) - f^{Y(0)}_{D=0,t=0}(y), \text{ for all } y. \tag{25}$$

To complete the proof, we show that (23) is equivalent to (25). Note that we can write

$$\begin{aligned}
f^{Y(0)}_{D=1,t=1}(y) &= \mathbb{E}_{P_D}\left[\frac{1}{N_1}\sum_i D_i 1[Y_{i,t=1}(0) = y]\right] \\
&= \frac{1}{N_1}\sum_i \pi_i 1[Y_{i,t=1}(0) = y] \\
&= \frac{1}{N_1}\sum_i \left(\dot{\pi}_i + \frac{N_1}{N}\right) 1[Y_{i,t=1}(0) = y] \\
&= \frac{1}{N_1}\dot{p}_1(y) + \frac{1}{N}\sum_i 1[Y_{i,t=1}(0) = y]
\end{aligned}$$

where the third line uses the definition of $\dot{\pi}_i$ to solve for π_i. Analogously, we have

$$f^{Y(0)}_{D=1,t=0}(y) = \frac{1}{N_1}\dot{p}_0(y) + \frac{1}{N}\sum_i 1[Y_{i,t=0}(0) = y]$$
$$f^{Y(0)}_{D=0,t=1}(y) = -\frac{1}{N_0}\dot{p}_1(y) + \frac{1}{N}\sum_i 1[Y_{i,t=1}(0) = y]$$
$$f^{Y(0)}_{D=0,t=0}(y) = -\frac{1}{N_0}\dot{p}_0(y) + \frac{1}{N}\sum_i 1[Y_{i,t=0}(0) = y]$$

Combining these results, we have that

$$(f^{Y(0)}_{D=1,t=1}(y) - f^{Y(0)}_{D=1,t=0}(y)) - (f^{Y(0)}_{D=0,t=1}(y) - f^{Y(0)}_{D=0,t=0}(y)) = \left(\frac{1}{N_1} + \frac{1}{N_0}\right)(\dot{p}_1(y) - \dot{p}_0(y)).$$

The result follows immediately.

□

Proof of Proposition E.2

Proof. Fix y. Let $g(\tilde{y}) = (\tilde{y} - 1)$ for $\tilde{y} \leqslant y$ and $g(\tilde{y}) = \tilde{y}$ for $\tilde{y} > y$. Observe that

$$\begin{aligned}
&\mathbb{E}_{P_D}\left[\hat{\tau}(\mathbf{D}, g(\boldsymbol{Y}))\right] \\
&= \tau_{ATT}(g(\mathbf{Y}(\bullet))) \\
&= \frac{1}{N}\sum_i \pi_i(g(Y_{i,t=1}(1)) - g(Y_{i,t=1}(0))) \\
&= \frac{1}{N}\sum_i \pi_i(Y_{i,t=1}(1) - Y_{i,t=1}(0)) - \frac{1}{N}\sum_i \pi_i 1[Y_{i,t=1}(1) \leqslant y] + \frac{1}{N}\sum_i \pi_i 1[Y_{i,t=1}(0) \leqslant y]. \quad (26)
\end{aligned}$$

Additionally, by assumption,

$$\mathbb{E}_{P_D}\left[\hat{\tau}(\mathbf{D}, \mathbf{Y})\right] = \tau_{ATT}(\mathbf{Y}(\bullet)) = \frac{1}{N}\sum_i \pi_i(Y_{i,t=1}(1) - Y_{i,t=1}(0)),$$

which is the first term in (26). Next, observe that

$$\mathbb{E}_{P_D}\left[\frac{1}{N}\sum_i D_i 1[Y_{i,t=1} \leqslant y]\right] = \frac{1}{N}\sum_i \pi_i 1[Y_{i,t=1}(1) \leqslant y],$$

which is the second term in (26). Finally, from the definition of $F^{Y(0)}_{D=1,t=1}$ we see that

$$F^{Y(0)}_{D=1,t=1}(y) = \mathbb{E}_{P_D}\left[\frac{1}{N}\sum_i D_i 1[Y_{i,t=1}(0) \leqslant y]\right] = \frac{1}{N}\sum_i \pi_i 1[Y_{i,t=1}(0) \leqslant y],$$

which is the third term in (26). Combining the results above, we have that

$$F_{D=1,t=1}^{Y(0)}(y) = \mathbb{E}_{P_D}\left[\frac{N}{N_1}\left[\hat{\tau}(\mathbf{D},\mathbf{Y}) - \hat{\tau}(\mathbf{D},g(\mathbf{Y})) - \frac{1}{N}\sum_i D_i 1[Y_{i,t=1} \leqslant y]\right]\right], \tag{27}$$

which gives the desired result. □

Appendix References

Armstrong, Timothy and Michal Kolesar, "Optimal Inference in a Class of Regression Models," *Econometrica*, 2018, *86*, 655–683.

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

_ and _ , "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *arXiv:1808.05293 [cs, econ, math, stat]*, August 2018.

Fisher, R. A., *The design of experiments* The design of experiments, Oxford, England: Oliver & Boyd, 1935. Pages: xi, 251.

Frandsen, Brigham R, "The Effects of Collective Bargaining Rights on Public Employee Compensation: Evidence from Teachers, Firefighters, and Police," *ILR Review*, 2016, *69* (1), 84–112.

Kolesar, Michal and Christoph Rothe, "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 2018, *108* (8), 2277–2304.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

Neyman, Jerzy, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1923, *5* (4), 465–472. Publisher: Institute of Mathematical Statistics.

Noack, Claudia and Christoph Rothe, "Bias-Aware Inference in Fuzzy Regression Discontinuity Designs," *arXiv:1906.04631 [econ.EM]*, 2020.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602 [econ, stat]*, August 2020. arXiv: 2008.00602.

— **and** — , "An Honest Approach to Parallel Trends," *Working Paper*, 2020.