CRESTCENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

WORKING PAPER SERIES

TWO-WAY FIXED EFFECTS AND DIFFERENCES-IN-DIFFERENCES IN HETEROGENEOUS ADOPTION DESIGNS WITHOUT STAYERS

Clément de Chaisemartin, Diego Ciccia, Xavier D’Haultfoeuille and Felix Knau

N°01 /January 2025

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

Two-way Fixed Effects and Differences-in-Differences in Heterogeneous Adoption Designs without Stayers*

Clément de Chaisemartin[†] Diego Ciccia[‡] Xavier D'Haultfœuille[§]
Felix Knau[¶]

First version: November 23, 2022
This version: July 26, 2024

Abstract

We consider treatment-effect estimation under a parallel trends assumption, in designs where no unit is treated at period one, all units receive a strictly positive dose at period two, and the dose varies across units. There are therefore no true control groups in such cases. First, we develop a test of the assumption that the treatment effect is mean independent of the treatment, under which the commonly-used two-way-fixed-effects estimator is consistent. When this test is rejected or lacks power, we propose alternative estimators, robust to heterogeneous effects. If there are units with a period-two treatment arbitrarily close to zero, the robust estimator is a difference-in-difference using units with a period-two treatment below a bandwidth as controls. Without such units, we propose non-parametric bounds, and an estimator relying on a parametric specification of treatment-effect heterogeneity. We use our results to revisit Pierce and Schott (2016) and Enikolopov et al. (2011).

*We are very grateful to Marc Gurgand for a discussion that initiated this paper. We are also grateful to Justin Pierce for his help accessing the data used in Pierce and Schott (2016). We are grateful to Tomasz Olma for his helpful comments. Clément de Chaisemartin was funded by the European Union (ERC, REALLYCREDIBLE,GA N°101043899).

[†]Sciences Po Paris, clement.dechaisemartin@sciencespo.fr
[‡]Northwestern University, Kellogg School of Management, diego.ciccia@kellogg.northwestern.edu
[§]CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr
[¶]University of Munich, felix.knau@students.uni-mannheim.de.

1 Introduction

We consider treatment-effect estimation with a group-level (e.g. county-level) panel data, in heterogeneous adoption designs (HAD) without stayers. In such designs, groups are untreated in a pre-period, henceforth period one, and receive a strictly positive treatment dose in a post-period, henceforth period two, and the dose varies across units.

Such designs are common. They arise when a policy is implemented universally, but exposure to the policy varies. For instance, the creation of Medicare Part D affected drugs differentially, depending on their Medicare market share (Duggan and Scott Morton, 2010). Similarly, from 2011 to 2013 the UK government significantly cut several welfare programs, but those cuts affected districts differentially, depending on their prior usage of those programs (Fetzer, 2019). Similarly, China's accession to the WTO eliminated uncertainty on US-China trade tariffs, but this affected US industries differentially, depending on their prior levels of tariffs' uncertainty (Pierce and Schott, 2016). Such designs also arise when an innovation is introduced, with an exposure rate that varies across sectors or over space. For instance, in 1996, a new TV channel was introduced in Russia. At that time, it was the only TV channel not controlled by the government, so one may be interested in the effect of having access to this independent news source on voting behavior. Exposure to this channel was heterogeneous across regions, due to varying signal quality across the Russian territory: in some regions, close to 100% of the population had access to it, while that rate was lower in other regions (Enikolopov et al., 2011).

Importantly, in all these examples, there are no stayers who remain unaffected in the post period: all drugs were affected by the introduction of Medicare Part D, all districts experienced some budgetary cuts during the UK austerity, tariffs' uncertainty was reduced in all US industries when China joined the WTO, and a strictly positive fraction of the population of all Russian regions gained access to the new TV channel in 1996.

Despite the absence of a true control group, a common strategy to estimate the treatment's effect in those designs is to run a so-called two-way fixed effects (TWFE) regression, as one would do in a classical difference-in-difference design. One regresses $Y_{g,t}$, the outcome of group g at period t, on group fixed effects, an indicator for period two, and $D_{g,t}$, the treatment dose of g at t. As $D_{g,1} = 0$, the coefficient on $D_{g,t}$ in this regression is equal to that on $D_{g,2}$ in a regression of ΔY_g on a constant and $D_{g,2}$, where $\Delta Y_g = Y_{g,2} - Y_{g,1}$

denotes g's outcome change from period one to two.

In previous work, we had already considered HADs,[1] and had shown a negative result concerning the TWFE estimator. Specifically, Theorem S1 of de Chaisemartin and D'Haultfœuille (2015) and Proposition S1 of de Chaisemartin and D'Haultfœuille (2020) show that in HADs, TWFE regressions may compare the outcome evolutions of "strongly" and "weakly" treated groups. The treatment effects of weakly treated groups get differenced out in those comparisons, and accordingly, TWFE regressions may fail to identify a convex combination of treatment effects, and could be misleading if treatment effects are heterogeneous. In HADs where there are stayers, namely groups that stay untreated at period two, difference-in-difference (DID) estimators robust to heterogeneous treatment effects have already been proposed (de Chaisemartin and D'Haultfœuille, 2018, 2020, 2021; Callaway et al., 2021). Those estimators compare the outcome evolution of groups that become treated to the outcome evolution of stayers, thus avoiding the forbidden comparisons leveraged by TWFE regressions. However, those estimators cannot be used in HADs without stayers, hence our focus on those common designs.

We start by showing a more positive result concerning the TWFE estimator. It estimates a well-defined effect if on top of a parallel trends assumption, one also assumes that treatment effects are mean-independent of the treatment variable. Under parallel-trends, this mean-independent-effects assumption has a testable implication:

$$E(\Delta Y_g|D_{g,2}) = \beta_0 + \beta_{fe} D_{g,2}, \tag{1}$$

namely the conditional mean of the outcome's evolution is linear in the period-two dose. We also show that under the parallel-trend assumption, there is actually an "if and only if" relationship between (1) and the mean-independent-effects assumption, in designs where we have quasi-stayers, namely groups such that $D_{g,2}$ is local to 0. If data is available for a period 0 before period 1, the parallel-trends assumption can be placebo tested by regressing $Y_{g,1} - Y_{g,0}$ on $D_{g,1}$, a pre-trends test often conducted by practitioners. This motivates the following procedure in designs with quasi-stayers: run a pre-trends test of the parallel-trends assumption, and run a test of (1); if neither test is rejected, use the TWFE estimator. It follows from de Chaisemartin and D'Haultfoeuille (2024) that if parallel-trends and (1) hold, inference on β_{fe} conditional on not rejecting the two pre-tests

[1]We coined the HAD terminology in Assumption S2 of de Chaisemartin and D'Haultfœuille (2020).

can at worst be conservative, but can never be liberal. As the interpretation of our linearity test depends on whether or not there are quasi-stayers, we also propose a non-parametric and tuning-parameter-free test of the null hypothesis that there are quasi-stayers.

Equality (1) is testable whenever $D_{g,2}$ can take at least three values. When $D_{g,2}$ can take a finite number of values K with $K \geq 3$, testing (1) is not difficult: one can just regress ΔY_g on a constant, $D_{g,2}$, $D_{g,2}^2$, ..., $D_{g,2}^{K-1}$, and run an F-test that the coefficients on $D_{g,2}^2$, ..., $D_{g,2}^{K-1}$ are all equal to zero. When $D_{g,2}$ can take an infinite number of values, testing (1) is more difficult. We use a test proposed by Stute (1997) and Stute et al. (1998). That test is non-parametric, tuning-parameter free, consistent, and it has power against local alternatives. For the purpose of this paper, we implemented the Stute test in the `stute_test` Stata and R packages. Those packages might be of interest to researchers interested in conducting specification tests outside of HADs. The Stute test, which relies on the wild bootstrap, is impractical in large datasets of more than 100,000 groups. Then, we propose another non-parametric tuning-parameter-free test, which generalizes a test proposed by Yatchew (1997) by allowing for heteroscedasticity. This test, implemented for the purposes of this paper in the `yatchew_test` Stata and R packages, does not rely on the bootstrap and can be computed in less than a minute in datasets with as many as 50,000,000 groups. We show that our heteroscedasticity-robust Yatchew test is less powerful than the Stute test, so we recommend using it only with very large datasets, where power is less of a concern.

When a pre-trends test is not rejected but the test of (1) is rejected, or the test is not rejected but one worries it may lack power, we propose alternatives to the TWFE estimator, which rely on the same parallel-trends assumption but are robust to heterogeneous effects. We distinguish two cases.

First, in designs with some quasi-stayers, our heterogeneity-robust estimator uses groups with a period-two treatment below a bandwidth as a control group. We leverage results from the regression discontinuity designs (RDDs) and non-parametric estimation literature (see Imbens and Kalyanaraman, 2012; Calonico et al., 2014, 2018), to propose an optimal bandwidth and a robust confidence interval accounting for the estimator's first-order bias. Our estimator converges at the standard univariate non-parametric rate, whereas heterogeneity-robust estimators previously proposed for designs with stayers converge at the standard parametric rate. Hence, moving from a design with stayers to a design with quasi-stayers comes with a precision cost. Our estimator and its confidence interval are computed by the `did_had` Stata and R packages. Those packages heavily rely on the `nprobust` Stata and R packages (Calonico et al., 2019), which should be cited, together with Calonico et al. (2018), whenever `did_had` is used.

Second, in designs without quasi-stayers, we propose two approaches. First, we propose non-parametric bounds under a boundedness condition on the treatment effect's magnitude. The sharp bounds are intersection bounds, so we can rely on Chernozhukov et al. (2013) and on their companion `clrbound` Stata package (Chernozhukov et al., 2015) for estimation and inference. Second, we propose an estimator of the average treatment effect under a parametric functional-form assumption on treatment-effect heterogeneity.

We use our results to revisit Pierce and Schott (2016) and Enikolopov et al. (2011). In Pierce and Schott (2016), an application where we do not reject the null that there are quasi-stayers, we find that after accounting for group-specific linear trends, the pre-trends test is not rejected, and the test of (1) is also not rejected. Then, the TWFE estimator with group-specific linear trends may be reliable in this application. Using that estimator, we find negative effects of China's WTO accession on US manufacturing employment, though the effects we find are slightly smaller than those in the original paper. As this application has a small sample size, our test of (1) may lack power, so as a robustness check we also compute our estimator under a parametric functional-form assumption on treatment-effect heterogeneity, allowing again for group-specific linear trends. Doing so yields similar results as with the TWFE estimator. Turning to our second application, in Enikolopov et al. (2011) we reject the null that there are quasi-stayers. We cannot test the parallel trends assumption as we do not have two time periods before the treatment's introduction. Our test of (1) is rejected, so the TWFE estimator may not be reliable. Our non-parametric bounds are wide and uninformative, even under tight bounds on the treatment's effect. Our estimator under a parametric functional-form assumption on treatment effects is much more noisy than, and sometimes very different from, the TWFE estimator. Thus, it seems more difficult to be conclusive on causal effects in this application.

Related Literature. Our paper is related to the recent heterogeneity-robust DID literature, and in particular to de Chaisemartin and D'Haultfœuille (2018, 2020, 2021) and Callaway et al. (2021), who also consider HADs. Unlike those papers, our paper's focus is on designs without stayers. In a conference proceedings posterior to this paper's first version, de Chaisemartin et al. (2024) also considered designs without stayers, but in the case where the treatment varies at baseline, rather than being equal to zero for all units. The most recent version of Callaway et al. (2021), which is posterior to this paper's first version, briefly discusses designs without stayers, but it does not propose treatment-effect estimators without stayers, and it does not distinguish between cases with and without quasi-stayers. Our paper also borrows from the literatures on functional-form specification tests (Stute, 1997; Stute et al., 1998), RDDs and non-parametric estimation (Imbens and Kalyanaraman, 2012; Calonico et al., 2014, 2018), and intersection bounds (Chernozhukov et al., 2013).

Organization of the paper. In Section 2, we present the set-up, introduce notation and discuss our main assumptions. In Section 3, we study the TWFE estimator in heterogeneous adoption designs and introduce our test of (1). In Section 4, we introduce our heterogeneity-robust estimators. In Section 5 we discuss how our results extend to applications with several time periods. In Section 6, we revisit two empirical applications. Most proofs appear in the text, but some longer proofs are collected in the appendix.

2 Setup, assumptions, and target parameters

2.1 Setup and assumptions

Group-level panel data. We are interested in estimating the effect of a treatment on an outcome, using a group-level (e.g. county-level, state-level) panel data set,[2] with G groups and two time periods. We discuss below how results generalize to designs with more time periods. Henceforth, groups are indexed by g and time periods by t.

Treatment and potential outcomes. Let $D_{g,t}$ denote the treatment of group g at t. The potential outcome of group g at t under treatment d is $Y_{g,t}(d)$, and the observed outcome is $Y_{g,t} := Y_{g,t}(D_{g,t})$. Our potential outcome notation rules out anticipatory effects: groups' outcome at period one does not depend on their period-two treatment. Our potential outcome notation also rules out dynamic or carry-over effects: groups' outcome at

[2] Of course, groups could actually be individuals or firms. Also, the estimators we consider below are not weighted by groups' populations, but considering weighted estimators is a mechanical extension.

period two does not depend on their past treatments. In the designs we consider, groups are all untreated at period one, so with two time periods, the no carry-over effects assumption is not of essence, we just impose it to simplify notation. On the other hand, the no-anticipation assumption is of essence.

Independent and identically distributed sample.

Assumption 1 (i.i.d. sample) $(Y_{g,1}, Y_{g,2}, D_{g,1}, D_{g,2})_{g=1,...,G}$ *are i.i.d.*

As groups are i.i.d., we drop the g subscript below, except when we introduce estimators.

Heterogeneous adoption designs without stayers. We consider heterogeneous adoption designs (HAD), where all groups are untreated at period one, and treated groups receive heterogeneous treatment doses at period two. Previous papers have considered such designs, assuming that there are stayers, namely groups that stay untreated at period two (de Chaisemartin and D'Haultfœuille, 2020, 2021; Callaway et al., 2021). Instead, in this paper we assume that all groups receive a strictly positive dose of treatment at period two, with variability in the dose received.

Design 1 *(HAD without stayers)* $D_1 = 0$*,* $D_2 > 0$*, and* $V(D_2) > 0$*.*

Our results also apply to designs where groups all receive the same non-zero treatment dose $d_1 \neq 0$ at period one. What is key is that all groups receive the same period-one dose.

HAD without stayers but with some quasi-stayers. Some of our results apply to a subset of Design 1, namely designs with quasi-stayers.

Design 1' *(HAD without stayers but with some quasi-stayers) The conditions in Design 1 hold and the support of* D_2 *includes 0.*[3]

The condition in Design 1' holds when there are no stayers ($P(D_2 = 0) = 0$), but there are groups whose period-two treatment is "very close" to zero: for any $\delta > 0$, $P(0 < D_2 < \delta) > 0$. For instance, this condition holds if D_2 is continuously distributed on $\mathbb{R}_+$ with a continuous density that is strictly positive at 0. Hereafter, groups with a period-two treatment close to zero are referred to as quasi-stayers.

[3]Recall that the support of a random variable A is the smallest closed set C such that $P(A \in C) = 1$.

Example 1: the effect on US employment of reducing uncertainty on tariffs with China. Since 1980, United States (US) imports from China have been subject to the low Normal Trade Relations (NTR) tariff rates reserved to WTO members. However, those rates required uncertain and politically contentious annual renewals. Without renewal, US import tariffs on Chinese goods would have spiked to higher non-NTR tariff rates. When China joined the WTO in 2001, the US granted it Permanent NTR (PNTR). The reform eliminated a potential tariff spike, equal to the difference between the non-NTR and NTR tariff rates, referred to as the NTR gap. This NTR gap varies substantially across industries: without NTR renewal, in some industries there would have been a large increase in tariffs on Chinese imports, while in other industries the increase would have been smaller. Pierce and Schott (2016) study the effect of the NTR-gap treatment on US manufacturing employment. They define their treatment $D_{g,t}$ as the interaction of industry g's NTR gap and t being after 2001. The NTR gap is strictly positive in all industries. Therefore, letting for instance $t = 1$ denote year 2000 and $t = 2$ denote year 2001, $D_{g,1} = 0$, $D_{g,2} > 0$, and there is variability in the NTR gap of treated industries: the conditions in Design 1 are met. Pierce and Schott (2016) estimate the NTR-gap's density, and their Figure 2 shows that the estimated density is strictly positive at 0. This suggests that there are quasi-stayers in this application, something we will formally test later.

Example 2: the effect of having access to independent information on voting behavior in Russia. In 1996, a new TV channel called NTV was introduced in Russia. At that time, it was the only TV channel not controlled by the government. Enikolopov et al. (2011) study the effect of having access to this independent news source, using voting outcomes for 1,938 Russian subregions in the 1995 and 1999 elections. After 1996, NTV's coverage rate is strictly positive and heterogeneous across regions: while a large fraction of the population receives it in urbanized regions, a smaller fraction receives it in more rural regions. Yet, in the region with the lowest exposure rate to NTV, this rate is still equal to 28%. This suggests that there are no quasi-stayers in this application, something we will formally test later. The authors define their treatment as $D_{g,t}$, the proportion of the population having access to NTV in region g and year t, hereafter referred to as the NTV exposure rate. We have $D_{g,1} = 0$, $D_{g,2} > 0$, and there is variability in the NTV exposure rate across treated regions: the conditions in Design 1 are met.

Identifying assumption. Throughout the paper, we maintain the following assumption.

Assumption 2 *(Strong exogeneity for the untreated outcome) There is a real number* μ_0 *such that* $E\left[\Delta Y(0)|D_2\right] = \mu_0$.

Assumption 2 requires that groups' untreated outcome evolution be mean independent of their period-two treatment: in the absence of treatment, groups receiving large and low treatment doses would not have experienced systematically different outcome trends. This assumption generalizes the standard parallel-trends assumption in classical DID designs to the more complex designs with a non-binary treatment dose we consider here. This assumption is also similar to a strong exogeneity assumption in panel data models.

Pre-trends test of Assumption 2. If the data contains another pre-period $t = 0$ where groups are all untreated, Assumption 2 can be "placebo-tested", for instance by regressing $Y_{g,1} - Y_{g,0}$ on $D_{g,2}$, because $Y_{g,1} - Y_{g,0} = Y_{g,1}(0) - Y_{g,0}(0)$ is an outcome evolution without treatment. We view the possibility of placebo-testing it as an important feature of Assumption 2: placebo tests are a crucial step in supporting the credibility of an identifying assumption in observational studies (Imbens et al., 2001; Imbens and Xu, 2024).

Exogeneity in levels? Instead of Assumption 2, one could assume

$$E\left[Y_2(0)|D_2\right] = \mu_0, \tag{2}$$

meaning that groups' period-two untreated outcome is mean independent of their period-two treatment. We believe that Assumption 2 is often more likely to hold than (2). (2) is also placebo testable, and in the two applications we revisit placebo tests of (2) are strongly rejected. However, if a researcher prefers to work under (2), they can apply the results below, replacing ΔY by Y_2.

2.2 Target parameters

Actual-versus-no-treatment slopes. As $D_2 > 0$, let

$$\mathrm{TE}_2 := \frac{Y_2(D_2) - Y_2(0)}{D_2}$$

denote the slopes of groups' potential outcome functions between 0 and their actual treatments, which we refer to as the actual-versus-no-treatment slope. As it is a slope, this effect can be interpreted as an effect of increasing the treatment by one unit. For a group receiving two doses of treatment at period two ($D_2 = 2$), $\text{TE}_2 = (Y_2(2) - Y_2(0))/2$. Of course, one may be interested in slopes of that group's potential outcome function at different treatment doses, such as $(Y_2(3) - Y_2(0))/3$. For $d_2 \neq D_2$, we refer to $(Y_2(d_2) - Y_2(0))/d_2$ as a counterfactual-versus-no-treatment slope. Estimating counterfactual-versus-no-treatment slopes requires estimating two unobserved potential outcomes, $Y_2(d_2)$ and $Y_2(0)$, while estimating actual-versus-no-treatment slopes only requires estimating one, $Y_2(0)$. While estimating $Y_2(0)$ may be achieved under a placebo-testable parallel trends assumption, estimating $Y_2(d_2)$ for $d_2 \neq D_2$ requires making stronger and non-placebo testable assumptions. This is the reason why we focus on actual-versus-no-treatment slopes, rather than on counterfactual-versus-no-treatment slopes. Another type of slopes one might be interested in is $(Y_2(D_2) - Y_2(d))/(D_2 - d)$ for $d > 0$, which we refer to as an actual-versus-counterfactual slope. When d tends to D_2, and assuming that $Y_2(d)$ is almost surely differentiable everywhere, this slope converges towards $Y_2'(D_2)$, the derivative of the potential outcome function evaluated at D_2. $E(Y_2'(D_2))$ is called the average marginal effect, a parameter that has often been studied in the literature. Like TE_2, estimating an actual-versus-counterfactual slope only requires estimating one unobserved outcome. However, the unobserved outcome in the actual-versus-counterfactual slope, $Y_2(d)$, is not observed at $t = 1$ and in prior periods. Therefore, estimating actual-versus-counterfactual slopes requires making non-placebo testable assumptions. This is again why we focus on actual-versus-no-treatment slopes, rather than on actual-versus-counterfactual slopes or $Y_2'(D_2)$.

Averages of slopes. TE_2 is a group-specific effect that applies to only one group and cannot be consistently estimated under Assumption 1. We therefore turn attention to averages of the TE_2 slopes, that can be consistently estimated. We consider two averages of slopes:

$$\begin{aligned} \text{AS} :=& E\left(\text{TE}_2\right] \\ \text{WAS} :=& E\left[\frac{D_2}{E[D_2]}\text{TE}_2\right]. \end{aligned}$$

AS is the Average Slope of treated groups. This parameter generalizes the well-known average treatment effect on the treated parameter to our setting with a non-binary treatment. WAS is a weighted average of treated groups' slopes, where groups with a larger period-two treatment receive more weight. While WAS may look like a less natural target parameter than AS, de Chaisemartin et al. (2022) put forward an economic and a statistical argument to consider that parameter. First, they show that WAS is actually the relevant quantity to consider in a cost-benefit analysis. Second, estimating AS may sometimes be more difficult than estimating WAS. When there are quasi-stayers, namely treated groups with a value of D_2 close to zero, the denominator of TE_2 is close to zero for those groups. Then, estimators of those groups' slopes may suffer from a small-denominator problem, which could substantially increase their variance, thus making it impossible to estimate the AS at the standard $\sqrt{G}-$rate (see Graham and Powell, 2012; Sasaki and Ura, 2021). On the other hand,

$$\text{WAS} = E\left[\frac{D_2}{E[D_2]}\frac{Y_2(D_2) - Y_2(0)}{D_2}\right] = \frac{E[Y_2(D_2) - Y_2(0)]}{E[D_2]}, \tag{3}$$

so estimators of WAS are not affected by a small-denominator problem, even if there are treated groups with a value of D_2 close to zero.

3 TWFE estimator

3.1 Consistency under parallel trends and homogeneous and linear effects

Decomposition of the TWFE estimand. Let $\widehat{\beta}_{fe}$ denote the coefficient on $D_{g,2}$ in a regression of ΔY_g on a constant and $D_{g,2}$. $\widehat{\beta}_{fe}$ is equal to the coefficient on $D_{g,t}$ in a regression of $Y_{g,t}$ on group fixed effects, an indicator for period 2, and $D_{g,t}$. Let

$$\beta_{fe} := \frac{E((D_2 - E(D_2))\Delta Y)}{E((D_2 - E(D_2))D_2)}$$

denote the probability limit of $\widehat{\beta}_{fe}$ when $G \to +\infty$ under Assumption 1 and if $E[D_2^2] < \infty$ and $E[\Delta Y^2] < \infty$. As in Design 1,

$$\Delta Y = Y_2(D_2) - Y_1(0) = Y_2(D_2) - Y_2(0) + \Delta Y(0) = D_2\text{TE}_2 + \Delta Y(0),$$

one can show that under Assumption 2,

$$\beta_{fe} = E\left(\frac{(D_2 - E(D_2))D_2}{E\left((D_2 - E(D_2))D_2\right)}E(\text{TE}_2|D_2)\right). \tag{4}$$

Equation (4) is not a new result: it is essentially an asymptotic version of Proposition S1 in de Chaisemartin and D'Haultfœuille (2020), who also coined the HAD terminology. It says that β_{fe} is a weighted sum of the conditional average slopes (CAS) $E(\text{TE}_2|D_2 = d_2)$, across all treated groups in the super population, where $E(\text{TE}_2|D_2 = d_2)$ receives a weight proportional to $(d_2 - E(D_2))d_2$. Therefore, $\widehat{\beta}_{fe}$ is not consistent for AS in general. $\widehat{\beta}_{fe}$ may not even converge to a convex combination of CAS: some weights in (4) are negative if $P(0 < D_2 < E(D_2)) > 0$, as is for instance necessarily the case if there are no stayers.

$\widehat{\beta}_{fe}$ is consistent for AS if the CAS do not vary with D_2. If

$$E(\text{TE}_2|D_2) = \text{AS}, \tag{5}$$

then

$$\begin{aligned}
&E\left(\frac{(D_2 - E(D_2))D_2}{E\left((D_2 - E(D_2))D_2\right)}E(\text{TE}_2|D_2)\right)\\
=&\text{AS} \times E\left(\frac{(D_2 - E(D_2))D_2}{E\left((D_2 - E(D_2))D_2\right)}\right)\\
=&\text{AS}.
\end{aligned}$$

Condition (5) is sufficient but not necessary to have AS $= \beta_{fe}$. We have AS $= \beta_{fe}$ if and only if

$$\text{Cov}\left(\frac{(D_2 - E(D_2))D_2}{E\left((D_2 - E(D_2))D_2\right)}, E(\text{TE}_2|D_2)\right) = 0, \tag{6}$$

a condition weaker than (5) (see Corollary 1 in de Chaisemartin and D'Haultfœuille, 2020).

Connection between (5) and homogeneous and linear effect assumptions. Two sufficient conditions for (5) to hold are

$$E\left(\frac{Y_2(d_2) - Y_2(0)}{d_2}\middle| D_2 = d_2\right) = E\left(\frac{Y_2(d_2) - Y_2(0)}{d_2}\right), \tag{7}$$

$$E\left(\frac{Y_2(d_2) - Y_2(0)}{d_2}\right) = \text{AS}. \tag{8}$$

(7) is an homogeneous effect assumption, which assumes that the effect of moving treatment from 0 to d_2 is the same for groups with different treatment doses. (8) is a linear effect assumption. To ease exposition, we sometimes refer to (5) as an homogeneous and linear effect assumption, though (7) and (8) are sufficient but not necessary for (5).

3.2 The homogeneous and linear effect assumption is testable

Let $\beta_0 = E(\Delta Y) - \beta_{fe}E(D_2)$ denote the intercept of the TWFE regression.

Theorem 1 *Suppose that Assumption 2 holds.*

1. *In Design 1, if* (5) *holds, then* $E(\Delta Y|D_2) = \beta_0 + \beta_{fe}D_2$.

2. *In Design 1', if* $E(\Delta Y|D_2) = \beta_0 + \beta_{fe}D_2$, *then* (5) *holds.*

Proof of Theorem 1. In Design 1,

$$
\begin{aligned}
E(\Delta Y|D_2) =& E(\Delta Y(0)|D_2) + D_2E(\mathrm{TE}_2|D_2)\\
=& \mu_0 + D_2E(\mathrm{TE}_2|D_2),
\end{aligned}
\tag{9}
$$

where the second equality follows from Assumption 2. Point 1 of Theorem 1 directly follows from plugging (5) into (9) and from the fact that if $E(U|V) = a_0 + a_1V$ then it is equal to the linear regression of U on $(1, V)$.

Then, assume that

$$E(\Delta Y|D_2) = \beta_0 + \beta_{fe}D_2. \tag{10}$$

Because $0 \in \mathrm{Supp}(D_2)$ in Design 1' and since the right-hand side of (10) is continuous, $E(\Delta Y|D_2 = 0)$ is well-defined. Then, equating (9) and (10) at $D_2 = 0$ yields $\beta_0 = \mu_0$. Further, equating (9) and (10) implies that

$$D_2E(\mathrm{TE}_2|D_2) =\beta_{fe}D_2,$$

and dividing by $D_2 > 0$ yields $E(\mathrm{TE}_2|D_2) = \beta_{fe}$. Taking expectations on both sides finally yields β_{fe} =AS. This proves Point 2 of Theorem 1 **QED.**

Interpreting a rejection of the linearity of $E(\Delta Y|D_2)$. Point 1 of Theorem 1 shows that under Assumption 2, if (5) holds, then $E(\Delta Y|D_2)$ is linear. By contraposition, if $E(\Delta Y|D_2)$ is not linear, then (5) cannot hold. This may suggest that $\widehat{\beta}_{fe}$ is not consistent for the AS, with the caveat that the homogeneous-and-linear-effect assumption in (5) is sufficient but not necessary to have that $\widehat{\beta}_{fe}$ is consistent for the AS. Unfortunately, it is either difficult or impossible to test (6), the necessary and sufficient condition for $\widehat{\beta}_{fe}$ to consistently estimate the AS. Specifically, with quasi-stayers, $E(\mathrm{TE}_2|D_2 = d_2)$ can be consistently estimated by DID estimators comparing the outcome evolutions of treated units with dose d_2 to the outcome evolution of quasi-stayers. However, as will become clear later, the covariance in (6) can only be estimated at the $G^{2/5}$ rate, so a test of (6) will lack power. Without quasi-stayers, under Assumption 2 alone testing (6) is not possible, as then $E(\text{TE}_2|D_2 = d_2)$ is not identified.

Interpreting a failure to reject the linearity of $E(\Delta Y|D_2)$, in designs with quasi-stayers. In designs with quasi-stayers, Point 2 of Theorem 1 shows that under Assumption 2, there is an "if and only if" relationship between the homogeneous-and-linear-effect assumption in (5) and the linearity of $E(\Delta Y|D_2)$. Therefore, if $E(\Delta Y|D_2)$ is linear then the homogeneous-and-linear-effect assumption holds, thus implying that $\widehat{\beta}_{fe}$ is consistent for the AS. This suggests that in designs with quasi-stayers, when a linearity test of $E(\Delta Y|D_2)$ and a pre-trends test of Assumption 2 are not rejected, one may use $\widehat{\beta}_{fe}$. As discussed below, under the null of linearity, the pre-test of the linearity of $E(\Delta Y|D_2)$ does not lead to liberal inference on β_{fe} conditional on not rejecting the test.

Interpreting a failure to reject the linearity of $E(\Delta Y|D_2)$, in designs without quasi-stayers. If there are no quasi-stayers, we no longer have an "if and only if" between (5) and $E(\Delta Y|D_2) = \beta_0 + \beta_{fe} D_2$: $E(\Delta Y|D_2) = \beta_0 + \beta_{fe} D_2$ could hold but $E(TE_2|D_2) = \beta_{fe} + (\beta_0 - \mu_0)/D_2$. As the interpretation of linearity tests depends on whether there are quasi-stayers or not, in Section 3.4 we propose a test of the null that there are quasi-stayers. In one of our two empirical applications, that test is not rejected.

3.3 Testing the homogeneous and linear effect assumption

A non-parametric and tuning-parameter-free test of the linearity of $E(\Delta Y|D_2)$, when D_2 takes a finite number of values. Assume that D_2 takes K values. If $K = 2$, we necessarily have $E(\Delta Y|D_2) = \beta_0 + \beta_{fe} D_2$, so there is no room for testability. If $K > 2$, to test that $E(\Delta Y|D_2)$ is linear one can just regress ΔY_g on a constant, $D_{g,2}$, $D_{g,2}^2$, ... $D_{g,2}^{K-1}$, and test that the coefficients on $D_{g,2}^2$, ... $D_{g,2}^{K-1}$ are all zero.

A non-parametric and tuning-parameter-free test of the linearity of $E(\Delta Y|D_2)$, when D_2 is continuous. When D_2 is continuous, we rely on Stute (1997) to test the linearity of $E(\Delta Y|D_2)$. Under the null hypothesis that $E(\Delta Y|D_2)$ is linear, then $(\widehat{\varepsilon}_{\text{lin},g})_{g=1,\dots,G}$, the residuals of the linear regression of ΔY_g on $D_{g,2}$, should not be correlated with any function of D_2. Then, consider the so-called cusum process of the residuals:

$$c_G(d) := G^{-1/2} \sum_{g=1}^{G} \mathbb{1}\left\{D_{g,2} \leq d\right\} \widehat{\varepsilon}_{\text{lin},g}.$$

Stute (1997) shows that under the null hypothesis, c_G, as a process indexed by d, converges to a Gaussian process. On the other hand, under the alternative, $c_G(d)$ tends to infinity for some d. Then, one can consider Kolmogorov-Smirnov or Cramér-von Mises test statistics based on $c_G(d)$. We use a Cramér-von Mises test statistic:

$$S := \frac{1}{G} \sum_{g=1}^{G} c_G^2(D_{g,2}).$$

To help build intuition, notice that sorting the data by $D_{g,2}$ and denoting by (g) the resulting indexation, one has that

$$S = \sum_{g=1}^{G} \left(\frac{g}{G}\right)^2 \left(\frac{1}{g} \sum_{h=1}^{g} \hat{\varepsilon}_{lin,(h)}\right)^2.$$

Now, $(1/g)\sum_{h=1}^{g} \hat{\varepsilon}_{lin,(h)} \approx E(\varepsilon_{\text{lin}}|D_2 \leq D_{2,g})$, and the null hypothesis that $E(\varepsilon_{\text{lin}}|D_2) = 0$ holds if and only if $E(\varepsilon_{\text{lin}}|D_2 \leq d_2) = 0$ for all d_2 in the support of D_2. The limiting distribution of S under the null is complicated, but Stute et al. (1998) show that one can approximate it using the wild bootstrap. Specifically, consider i.i.d. random variables $(\eta_g)_{g=1,\dots,G}$ with $E[\eta_g] = 0$, $E[\eta_g^2] = E[\eta_g^3] = 1$.[4] Then, let $\widehat{\varepsilon}^*_{\text{lin},g} := \widehat{\varepsilon}_{\text{lin},g}\eta_g$ and

$$\Delta Y_g^* = \widehat{\beta}_0 + \Delta D_g \widehat{\beta}_{fe} + \widehat{\varepsilon}^*_{\text{lin},g}.$$

Then, we compute S^*, the bootstrap counterpart of S based on the sample $(D_g, \Delta Y_g^*)_{g=1,\dots,G}$.

Properties of the Stute test. Stute (1997) and Stute et al. (1998) show that the test has asymptotically correct size, is consistent under any fixed alternative, and has non-trivial power against local alternatives converging towards the null at the $1/G^{1/2}$ rate. Moreover, it follows from Corollary 1 in de Chaisemartin and D'Haultfoeuille (2024) that under the null, inference on β_{fe} conditional on not rejecting the Stute test is conservative. Thus, under the null of linearity, pre-testing cannot make post-test inference liberal.

[4] In practice, we use the standard two-point distribution: $\eta_g = (1+\sqrt{5})/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$, $\eta_g = (1-\sqrt{5})/2$ otherwise.

Computation of the Stute test. The `stute_test` Stata (see de Chaisemartin et al., 2024d) and R (see de Chaisemartin et al., 2024c) commands compute the Stute test. The test's p-value computation relies on the wild bootstrap, which is demanding in terms of computational power and speed. To reduce computation time, the Stata and R commands implementing the test use a vectorization of the test statistic, see Online Appendix B for further details. With this method, the test runs very quickly with moderate sample sizes. For instance, it takes less than one second with $G = 5,000$. It still runs in less than two minutes with $G = 50,000$. However, the vectorization requires specifying a $G \times G$ matrix, so starting at around $G = 100,000$ the test does not run anymore on standard computers, which cannot allocate enough memory to store such a large matrix.

Another non-parametric and tuning-parameter-free test of the linearity of $E(\Delta Y|D_2)$, for large datasets. For large datasets, we recommend another test that does not rely on the bootstrap, and whose computation time remains below one minute, even for datasets as large as $G = 50,000,000$. Yatchew (1997) proposed a tuning-parameter-free test of linearity. However, his test is not entirely non-parametric, as it assumes homoscedasticity. In Online Appendix C, we show that with heteroscedasticity, the Yatchew test is liberal, and we propose an heteroscedasticity-robust (HR) version of this test. We show that the test statistic converges towards a standard normal distribution under the null of linearity, so our HR Yatchew test does not need to rely on the bootstrap. In Theorem 5, we show that our HR Yatchew test has asymptotically correct size, is consistent under any fixed alternative, and that under the null of linearity the TWFE estimator and the test statistics are asymptotically independent so that pre-testing does not distort inference. On the other hand, we also show that our HR Yatchew test only has non-trivial power against local alternatives converging towards the null at the $1/G^{1/4}$ rate, which leads it to be less powerful than the Stute test. This is the reason why we only recommend our HR Yatchew test for large data sets: then computing the Stute test might be very slow or infeasible, while power is less of a concern with large data sets. We have developed the `yatchew_test` Stata and R commands to compute the original Yatchew test as well as our HR version. To our knowledge, the Stute and HR Yatchew tests are the only available linearity tests that are non-parametric and tuning-parameter-free.

The Stute and HR Yatchew tests can also be used to placebo-test the parallel trends condition in Assumption 2. Assumption 2 is a mean-independence condition. If the data contains another pre-period $t = 0$ where groups are all untreated, as in period $t = 1$, regressing $Y_{g,1} - Y_{g,0}$ on $D_{g,2}$, is a placebo test of a condition weaker than Assumption 2, namely that $\Delta Y(0)$ is uncorrelated with D_2. To placebo test Assumption 2, one should test the null that $E(Y_1 - Y_0|D_2)$ is constant, something that can also be achieved with a slightly different Stute or Yatchew test: in the definition of the test statistics, one just needs to replace the residuals from a linear regression of $Y_1 - Y_0$ on a constant and D_2 by the residuals from a linear regression of $Y_1 - Y_0$ on a constant.

When the treatment dose varies at baseline, the Stute and HR Yatchew tests are less useful to assess the validity of TWFE estimators. Assume one uses a two-period panel data set to estimate a treatment's effect, but $D_1 \neq 0$ and $V(D_1) > 0$, so the conditions in Design 1 are not met. Then, the TWFE estimator is the coefficient on ΔD_g in a regression of ΔY_g on ΔD_g. Letting $TE_t = (Y_t(D_t) - Y_t(0))/D_t$, $Y_t = Y_t(0) + D_t TE_t$. If one is ready to assume that the treatment effect is constant over time ($\text{TE}_2 = \text{TE}_1$), then

$$\Delta Y = \Delta Y(0) + \Delta D \times \text{TE}_2.$$

Then, one can show that under Assumption 2, if there are stayers or quasi-stayers there is an "if and only if" between

$$E(\Delta Y|\Delta D) = \beta_0 + \beta_{fe}\Delta D$$

and $E(\text{TE}_2|\Delta D) = E(\text{TE}_2|\Delta D \neq 0)$, a condition under which the TWFE estimator is consistent for $E(\text{TE}_2|\Delta D \neq 0)$. However, this only holds if $\text{TE}_2 = \text{TE}_1$. If the treatment effect varies over time, as is often likely to be the case, then one might have that $E(\Delta Y|\Delta D)$ is linear but the TWFE estimator is not consistent for a well-defined causal effect.

3.4 Testing the null that there are quasi-stayers.

As the interpretation of the linearity tests crucially depends on whether or not there are quasi-stayers, we now propose a test of the null hypothesis that $\underline{d} := \inf \text{Supp}(D_2) = 0$, against $\underline{d} > 0$, where $\text{Supp}(D_2)$ denotes the support of D_2. We consider the following simple and tuning-parameter-free test, of nominal level α. The test statistic is $T = D_{2,(1)}^2/(D_{2,(2)}^2 -$

$D^2_{2,(1)}$), where $D_{2,(1)} \leq ... \leq D_{2,(G)}$ denotes the order statistic of $(D_{2,g})_{g=1,...,G}$. The critical region is $W_\alpha := \{T > 1/\alpha - 1\}$. Intuitively, we reject the null if the distance between $D^2_{2,(1)}$ and 0 is more than $1/\alpha - 1$ times larger than the distance between $D^2_{2,(2)}$ and $D^2_{2,(1)}$. Taking squares of the order statistics, rather than the order statistics themselves, ensures that the test remains valid if the density of D_2 is equal to zero at $\underline{d}$. We show that this test has asymptotic size equal to α and nontrivial local power on a broad class of cdfs for D_2. Specifically, let $\mathcal{D}$ denotes the set of cdfs on the real line, and for $k \in \{1, 2\}$, $\overline{d} > \underline{d} \geq 0$, $m > 0$ and $M > 0$, let us consider

$$\mathcal{F}^{k,\underline{d},\overline{d}}_{m,M} := \Big\{F \in \mathcal{D} : F \text{ is } k \text{ times differentiable on } [\underline{d}, \overline{d}],\ F^{(0)}(\underline{d}) = ... = F^{(k-1)}(\underline{d}) = 0,$$
$$F^{(k)}(d) \geq m\ \forall d \in [\underline{d}, \overline{d}] \text{ and } |F^{(k)}(d_2) - F^{(k)}(d_1)| \leq M|d_2 - d_1|\ \forall (d_1, d_2) \in [\underline{d}, \overline{d}]^2\Big\}.$$

$\mathcal{F}^{1,\underline{d},\overline{d}}_{m,M}$ and $\mathcal{F}^{2,\underline{d},\overline{d}}_{m,M}$ are two sets of cdfs whose support has infimum equal to $\underline{d}$. The main difference between them is that cdfs in $\mathcal{F}^{1,\underline{d},\overline{d}}_{m,M}$ have a density bounded from below (by $m > 0$) in a neighborhood of $\underline{d}$, whereas cdfs in $\mathcal{F}^{2,\underline{d},\overline{d}}_{m,M}$ have a density equal to zero at $\underline{d}$ but a derivative of the density bounded from below in a neighborhood of $\underline{d}$. Hereafter, we denote probabilities by P_F instead of P, to emphasize that they depend on the cdf of D_2.

Theorem 2 *Fix $\alpha \in (0, 1)$, $\overline{d} > \underline{d} > 0$, $m > 0$, and $M > 0$. We have:*

1. *(Asymptotic size control)* $\limsup_{G\to\infty} \sup_{F\in\mathcal{F}^{1,0,\overline{d}}_{m,M}} P_F(W_\alpha) \leq \alpha$ *and* $\limsup_{G\to\infty} \sup_{F\in\mathcal{F}^{2,0,\overline{d}}_{m,M}} P_F(W_\alpha) = \alpha$.

2. *(Uniform consistency) For any* $k \in \{1, 2\}$, $\liminf_{G\to\infty} \inf_{F\in\mathcal{F}^{k,\underline{d},\overline{d}}_{m,M}} P_F(W_\alpha) = 1$.

3. *(Local power with a strictly positive density at $\underline{d}$) For all* $(\underline{d}_G)_{G\geq1}$ *satisfying* $\liminf G\underline{d}_G \geq 1$, $\liminf_{G\to\infty} \inf_{F\in\mathcal{F}^{1,\underline{d}_G,\overline{d}}_{m,M}} P_F(W_\alpha) > \alpha$.

4. *(Local power with a density equal to zero at $\underline{d}$) For all* $(\underline{d}_G)_{G\geq1}$ *satisfying* $\liminf G^{1/2}\underline{d}_G > 0$, $\liminf_{G\to\infty} \inf_{F\in\mathcal{F}^{2,\underline{d}_G,\overline{d}}_{m,M}} P_F(W_\alpha) > \alpha$.

Point 1 of Theorem 2 establishes the asymptotic validity of the test: the test is asymptotically conservative if $F \in \mathcal{F}^{1,0,\overline{d}}_{m,M}$, and asymptotically exact if $F \in \mathcal{F}^{2,0,\overline{d}}_{m,M}$. Point 2 shows that the test is consistent against fixed alternatives. Finally, Points 3 and 4 show that the test has power against local alternatives: if $F \in \mathcal{F}^{1,0,\overline{d}}_{m,M}$, the test has power against alternatives converging towards the null at the G rate, while if $F \in \mathcal{F}^{2,0,\overline{d}}_{m,M}$ the test has power against alternatives converging towards the null at the $G^{1/2}$ rate. In Point 3, the restriction that $\liminf G\underline{d}_G \geq 1$ (instead of $\liminf G\underline{d}_G > 0$), is due to the fact that the test is conservative under the null. If one is willing to assume that $F \in \mathcal{F}^{1,\underline{d},\overline{d}}_{m,M}$, meaning that D_2's density is strictly positive at the infimum of its support, then one can replace T by $D_{2,(1)}/(D_{2,(2)} - D_{2,(1)})$, to obtain an exact and more powerful test. However, unlike our test, this test becomes liberal if D_2's density is equal to zero at the infimum of its support.

4 Heterogeneity-robust estimators

When pre-trends tests of Assumption 2 are not rejected, there are at least three instances where one may prefer using an heterogeneity-robust estimator instead of the TWFE estimator. First, the test of the homogeneous-and-linear-effect assumption in (5) may be rejected. Second, even when the test is not rejected, one might worry that it lacks power. Third, even when the test is not rejected and one does not worry about its power, one may be in a design without quasi-stayers, in which case (5) could fail even when its testable implication holds. We now propose alternative heterogeneity-robust estimators, considering first designs with quasi-stayers, before turning to designs without quasi-stayers.

4.1 Designs with quasi-stayers

In this section, we focus on WAS. AS can also be identified and estimated with quasi-stayers, but as mentioned above, its estimator is often significantly less precise than that of WAS, and it sometimes even converges at a slower rate.

Identification of WAS with quasi-stayers.

Theorem 3 *Suppose that we are in Design 1' and Assumption 2 holds. Then,*

$$WAS = \frac{E[\Delta Y] - E[\Delta Y | D_2 = 0]}{E[D_2]}. \tag{11}$$

Theorem 3 shows that with quasi-stayers, WAS is identified by an estimand comparing the outcome evolution of treated and untreated groups, and scaling that comparison by the average treatment of treated groups. The estimand in Theorem 3 is a special case of a Wald-DID estimand using a control group with a stable exposure to the treatment, as recommended by de Chaisemartin and D'Haultfœuille (2018) in fuzzy designs.

Proof of Theorem 3. As mentioned above, $E[\Delta Y|D_2=0]$ is well-defined in Design 1'. Moreover,

$$\begin{aligned} & E[\Delta Y]-E[\Delta Y|D_2=0] \\ =& E[Y_2(D_2)-Y_1(0)]-E[\Delta Y(0)|D_2=0] \\ =& E[Y_2(D_2)-Y_2(0)]+E[\Delta Y(0)]-E[\Delta Y(0)|D_2=0] \\ =& E[Y_2(D_2)-Y_2(0)]. \end{aligned}$$

The first equality follows from the fact we are in Design 1'. The third equality follows from Assumption 2. The result follows from the previous display and (3) **QED.**

Estimators' definition. As $P(D_2=0)=0$, estimating $E[Y_2-Y_1|D_2=0]$ is not straightforward, as there are no stayers. We propose to use "quasi-stayers", namely observations with D_2 close to 0, as the control group. Formally, we rely on local linear regression, as in regression discontinuity designs (RDDs) and more generally in non-parametric estimation. We define the following estimators, indexed by the bandwidth h:

$$\widehat{\beta}_h^{qs} := \frac{\frac{1}{G}\sum_{g=1}^G \Delta Y_g - \widehat{\mu}_h}{\frac{1}{G}\sum_{g=1}^G D_{g,2}},$$

with $\widehat{\mu}_h$ the intercept in the local linear regression of ΔY_g on $D_{g,2}$, weighting observations by $k(D_{g,2}/h)/h$, for a kernel function k and a bandwidth $h>0$.

Estimators' asymptotic distribution. Let $m(d_2):=E(\Delta Y|D_2=d_2)$. One can derive the asymptotic behavior of $\widehat{\beta}_h^{qs}$ under the conditions below:

Assumption 3 (Regularity conditions)

1. *The cumulative distribution function of D_2 is differentiable at 0, with derivative denoted by $f_{D_2}(0)$. Moreover, $f_{D_2}(0)>0$.*

2. *m, defined on Supp(D_2), is twice differentiable at $d=0$.*

3. $\sigma^2(d) := V(\Delta Y|D_2 = d)$*, defined on Supp*$(D_2)$*, is continuous at 0 and* $\sigma^2(0) > 0$.

4. k *is bounded and has bounded support.*

5. As $G \to \infty$*, the bandwidth* h_G *satisfies* $h_G \to 0$ *and* $Gh_G \to \infty$.

Assumption 3 imposes standard regularity conditions on the distribution of D_2, the function m, the kernel function and the bandwidth. We also introduce the following notation. Let $\kappa_k := \int_0^\infty t^k k(t)dt$ for $k \in \mathbb{N}$ and

$$k^*(t) := \frac{\kappa_2 - \kappa_1 t}{\kappa_0\kappa_2 - \kappa_1^2}k(t),$$
$$C := \frac{\kappa_2^2 - \kappa_1\kappa_3}{\kappa_0\kappa_2 - \kappa_1^2}.$$

Since $\sum_{g=1}^G \Delta Y_g/G$ and $\sum_{g=1}^G D_{g,2}/G$ are root-G consistent, their randomness is negligible compared to that of $\widehat{\mu}_h$. Thus,

$$G^{2/5}\left(\widehat{\beta}^{qs}_{h_G} - \text{WAS}\right) = G^{2/5}\frac{\widehat{\mu}_{h_G} - m(0)}{E[D_{g,2}]} + o_P(1).$$

Then, following the exact same reasoning as in, say, Imbens and Kalyanaraman (2012), we obtain that under Assumptions 1-3,

$$\sqrt{Gh_G}\left(\widehat{\beta}^{qs}_{h_G} - \text{WAS} - h_G^2\frac{Cm''(0)}{2E[D_{g_2}(0)]}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(0)\int_0^\infty k^*(u)^2du}{E[D_{g_2}(0)]^2 f_{D_2}(0)}\right). \tag{12}$$

The fastest rate of convergence is obtained with $G^{1/5}h_G \to c > 0$, in which case

$$G^{2/5}\left(\widehat{\beta}^{qs}_{h_G} - \text{WAS}\right) \xrightarrow{d} \mathcal{N}\left(\frac{c^2Cm''(0)}{2E[D_{g_2}(0)]}, \frac{\sigma^2(0)\int_0^\infty k^*(u)^2du}{cE[D_{g_2}(0)]^2 f_{D_2}(0)}\right). \tag{13}$$

Optimal bandwidth and robust confidence interval. Based on (13), one can derive a so-called optimal bandwidth, which, as in RDDs (see Imbens and Kalyanaraman, 2012), minimizes the asymptotic mean squared error of $\widehat{\beta}^{qs}_{h_G}$. Then, inference on WAS is not straightforward, because the asymptotic distribution of

$$\sqrt{Gh_G^*}(\widehat{\beta}^{qs}_{h_G^*} - \text{WAS})$$

has a first-order bias that needs to be accounted for. However, the approach for local-polynomial regressions in Calonico et al. (2018) readily applies to our set up, in particular because it can be used to estimate a conditional expectation function at a boundary point. We rely on their results and software implementation (see Calonico et al., 2019) to:

1. estimate an optimal bandwidth $\widehat{h}^*_G$;

2. compute $\widehat{\mu}_{\widehat{h}^*_G}$;

3. compute $\widehat{M}_{\widehat{h}^*_G}$, an estimator of $\widehat{\mu}_{\widehat{h}^*_G}$'s first-order bias;

4. compute $\widehat{V}_{\widehat{h}^*_G}$, an estimator of the variance of $\widehat{\mu}_{\widehat{h}^*_G} - \widehat{M}_{\widehat{h}^*_G}$.

With those inputs, we can simply define our estimator with quasi-stayers as

$$\widehat{\beta}^{qs}_{\widehat{h}^*_G} := \frac{\frac{1}{G}\sum_{g=1}^G \Delta Y_g - \widehat{\mu}_{\widehat{h}^*_G}}{\frac{1}{G}\sum_{g=1}^G D_{g,2}},$$

and its bias-corrected confidence interval as

$$\left[\widehat{\beta}^{qs}_{\widehat{h}^*_G} + \frac{\widehat{M}_{\widehat{h}^*_G}}{\frac{1}{G}\sum_{g=1}^G D_{g,2}} \pm \frac{q_{1-\alpha/2}\sqrt{\widehat{V}_{\widehat{h}^*_G}/(G\widehat{h}^*_G)}}{\frac{1}{G}\sum_{g=1}^G D_{g,2}}\right], \tag{14}$$

where q_x denotes the quantile of order x of a standard normal. We refer to Calonico et al. (2018) for conditions ensuring the asymptotic validity of this confidence interval.

Computation. The `did_had` Stata (see de Chaisemartin et al., 2024b) and R (see de Chaisemartin et al., 2024a) commands compute $\widehat{\beta}^{qs}_{\widehat{h}^*_G}$ and its bias-corrected confidence interval. `did_had` heavily relies on the `nprobust` package of Calonico et al. (2019), which should be cited, together with Calonico et al. (2018), whenever `did_had` is used.

Simulations. Our bias-corrected confidence interval does not account for the variability of $\widehat{\beta}^{qs}_{\widehat{h}^*_G}$ stemming from $\sum_{g=1}^G \Delta Y_g/G$ and $\sum_{g=1}^G D_{g,2}/G$. That variability is asymptotically negligible with respect to that of $\widehat{\mu}_{\widehat{h}^*_G}$, but failing to account for it may lead to size distortions in finite samples. We conduct a small simulation study to assess whether this is a concern. In our simulations, D_2 follows a uniform distribution on $[0,1]$, $\Delta Y(0)$ follows a standard normal, and $m(d) = d + d^2$, thus implying that WAS $= 5/3$. Table 1 shows that with 100 groups, the 95% confidence interval slightly undercovers, while with 500 groups, the confidence interval has nearly nominal coverage. In both cases, $\widehat{\beta}^{qs}_{\widehat{h}^*_G}$ is slightly upward biased for WAS, and the bias correction makes it slightly downward biased. These results suggest that inference based on (14) should be reliable for moderately large sample sizes.

Table 1: Simulation Results

Sample size G	$\widehat{h}^*_G$	$\widehat{\beta}^{qs}_{\widehat{h}^*_G}$	Bias	Standard error	95% CI Coverage
100	0.258	1.701	0.042	1.229	0.907
500	0.256	1.673	0.023	0.543	0.941

Notes: "Bias" and "Standard error" are the true bias and standard error over the simulations.

4.2 Designs without quasi-stayers

Without quasi-stayers, $\mu_0 = E[\Delta Y(0)]$ and therefore AS and WAS are not point identified under Assumption 2. Then, we start by proposing bounds under non-parametric conditions, before imposing parametric conditions to recover point identification.

4.2.1 Partial identification

In this section, we propose bounds for WAS, under the assumption that the CAS are bounded. One could follow similar steps to propose bounds for AS.

Assumption 4 *There exists known $\underline{M} < \overline{M}$ such that almost surely,*

$$\underline{M} \leq E\left[TE_2|D_2\right] \leq \overline{M}.$$

Let $\underline{\mu} := \sup_{d_2 \in \text{Supp}(D_2)} m(d_2) - \overline{M}d_2$ and $\overline{\mu} := \inf_{d_2 \in \text{Supp}(D_2)} m(d_2) - \underline{M}d_2$.

Theorem 4 *Suppose that we are in Design 1.*

1. *If Assumptions 2 and 4 hold and $\underline{\mu} \leq \overline{\mu}$,*

$$B_- := \frac{E\left[Y_2 - Y_1 - \overline{\mu}\right]}{E\left[D_2\right]} \leq WAS \leq B_+ := \frac{E\left[Y_2 - Y_1 - \underline{\mu}\right]}{E\left[D_2\right]},$$

   and the bounds are sharp.

2. *If $\underline{\mu} > \overline{\mu}$, Assumptions 2 and 4 cannot jointly hold.*

A simple expression for the bounds in a special case. Let $\underline{d}_2 := \inf \text{Supp}(D_2)$. If

$$\underline{M} \leq \frac{\partial m}{\partial d_2}(d_2) \leq \overline{M} \quad \forall d_2 \in \text{Supp}(D_2),$$

then

$$B_- = \frac{E\left[Y_2 - Y_1\right] - \left(E[Y_2 - Y_1 | D_2 = \underline{d}_2] - \underline{M}\underline{d}_2\right)}{E\left[D_2\right]},$$

and

$$B_+ = \frac{E\left[Y_2 - Y_1\right] - \Big(E[Y_2 - Y_1 | D_2 = \underline{d}_2] - \overline{M}\underline{d}_2\Big)}{E\left[D_2\right]}.$$

The numerator of, say, B_- is a DID comparing treated groups to the least treated groups, with $D_2 = \underline{d}_2$, plus the term $\underline{M}\underline{d}_2$. That term accounts for the fact that the least treated groups are still treated, so their outcome evolution identifies groups' counterfactual outcome evolution without treatment, plus a bias term due to their treatment's effect, which can be bounded under Assumption 4. In that special case,

$$B_+ - B_- = (\overline{M} - \underline{M})\frac{\underline{d}_2}{E\left[D_2\right]}.$$

Accordingly, the lower the treatment dose received by the least-treated groups relative to the average dose received by treated groups, the more informative the bounds are.

Estimation and inference. We can rewrite the bounds as follows:

$$\begin{aligned} B_- &= \frac{E[Y_2 - Y_1]}{E[D_2]} + \frac{1}{E[D_2]} \sup_{d_2 \in \text{Supp}(D_2)} \theta_\ell(d_2), \\ B_+ &= \frac{E[Y_2 - Y_1]}{E[D_2]} + \frac{1}{E[D_2]} \inf_{d_2 \in \text{Supp}(D_2)} \theta_u(d_2), \end{aligned} \tag{15}$$

with $\theta_\ell(d_2) := E[\underline{M}D_2 + Y_1 - Y_2 | D_2 = d_2]$ and $\theta_u(d_2) := E[\overline{M}D_2 + Y_1 - Y_2 | D_2 = d_2]$. Thus, we can use Chernozhukov et al. (2013) to perform inference on WAS based on Theorem 4.

4.2.2 Point identification

Parametric assumption. In this section, we propose estimators of AS and WAS that can be used even if there are no stayers or quasi-stayers, under a parametric assumption.

Assumption 5 *There exist a known integer K, K known functions $f_1(d)$, ..., and $f_K(d)$ and K unknown real numbers δ_1, ..., δ_K such that*

$$E[TE_2 | D_2 = d] = \sum_{k=0}^{K} \delta_k f_k(d). \tag{16}$$

Assumption 5 is a parametric functional-form assumption on $E[\text{TE}_2|D_2 = d]$. It for instance holds if $E[\text{TE}_2|D_2 = d] = \delta_0 + \delta_1 d$, meaning that groups with different values of D_2 may have different CAS, but the relationship between CAS and D_2 is linear. Alternatively one could assume that $E[\text{TE}_2|D_2 = d]$ is a polynomial of order 2 or 3 in d. Under Assumption 5,

$$\text{AS} = E\left[\sum_{k=0}^{K} \delta_k f_k(D_2)\right],$$

and

$$\text{WAS} = E\left[\frac{D_2}{E(D_2)} \sum_{k=0}^{K} \delta_k f_k(D_2)\right],$$

so identifying $(\delta_0, ..., \delta_K)$ is sufficient to identify AS and WAS.

Identification and estimation of AS and WAS under Assumptions 2 and 5. Under Assumptions 2 and 5,

$$\begin{aligned} E[\Delta Y|D_2] =& E[\Delta Y(0)|D_2] + D_2 E[\text{TE}_2|D_2] \\ =& \mu_0 + \sum_{k=0}^{K} \delta_k D_2 f_k(D_2). \end{aligned} \tag{17}$$

It directly follows from (17) that $(\mu_0, \delta_0, ..., \delta_K)$ are the population coefficients from a regression of ΔY on $(1, D_2 f_0(D_2), D_2 f_1(D_2), ..., D_2 f_K(D_2))$. Therefore, AS and WAS are identified as soon as $(1, D_2 f_0(D_2), D_2 f_1(D_2), ..., D_2 f_K(D_2))$ are not linearly dependent. To estimate, say, AS, one can regress ΔY_g on $(1, D_{g,2} f_0(D_{g,2}), D_{g,2} f_1(D_{g,2}), ..., D_{g,2} f_K(D_{g,2}))$ and then use

$$\widehat{\beta}^{ns} := \frac{1}{G_1} \sum_{g: D_{g,2} > 0} \left(\sum_{k=0}^{K} \widehat{\delta}_k f_k(D_{g,2})\right).$$

For instance, if one assumes that $E[\text{TE}_2|D_2 = d] = \delta_0 + \delta_1 d$, to estimate AS one can just regress ΔY_g on $(1, D_{g,2}, D_{g,2}^2)$, and use $\widehat{\delta}_1 + \widehat{\delta}_2 \overline{D}_2$ as the AS estimator. One can follow similar steps to form an estimator $\widehat{\beta}_w^{ns}$ of WAS.

Testability of Assumption 5. Equation (17) shows that Assumption 5 is testable, by testing whether $E[\Delta Y|D_2]$ has the functional form therein. One could potentially use the Stute test to test (17). For instance, one could imagine a procedure where one starts by testing whether $E[\Delta Y|D_2]$ is linear. If the test is not rejected one estimates a simple TWFE regression, but if the test is rejected one tests whether $E[\Delta Y|D_2]$ is quadratic.

If the test is not rejected one regresses ΔY_g on $(1, D_{g,2}, D_{g,2}^2)$, but if the test is rejected one tests whether $E[\Delta Y|D_2]$ is cubic, etc. We do not recommend this procedure, because we conjecture that the Stute test becomes less powerful to detect departures from richer parametric functional forms. Also, while under the null of linearity pre-testing does not distort inference conditional on not rejecting the test, the aforementioned iterative pre-testing procedure does not have the same guarantees.

5 Extensions: more than two time periods.

No variation in treatment timing. Our results easily generalize to applications with several time periods, where treated groups all start receiving the treatment at the same time period F. Then, our results apply to every post-period $t \geq F$, letting periods 1 and 2 in the above respectively refer to periods $F-1$ and t. Accordingly, the Stute test can be used to assess the reliability of the coefficients from an event-study regression of $Y_{g,t}$ on period fixed effects, group fixed effects, and the treatment interacted with the period fixed effects, with period $F-1$ as the reference period. Indeed, one can show that the coefficients on the interactions are numerically equivalent to coefficients from regressions of $Y_{g,t} - Y_{g,F-1}$ on the treatment. The `stute_test` package can be used with several time periods and no variation in treatment timing. Then, to avoid multiple testing issues, it can compute a joint test that $E(Y_{g,t} - Y_{g,F-1}|D_{g,F})$ is linear in $D_{g,F}$ for all $t \geq F$. If the Stute test indicates that TWFE regressions may not be reliable, the `did_had` package can be used, and this package is also applicable in designs with several time periods and no variation in treatment timing.

Variation in treatment timing. On the other hand, the Stute test cannot be used to assess the reliability of TWFE coefficients in designs with variation in treatment timing. Moreover, note that in designs with variation in treatment timing, there must be some stayers, at least till the period where the last cohort gets treated. Then, the heterogeneity-robust DID estimators proposed in de Chaisemartin and D'Haultfœuille (2021) and computed by the `did_multiplegt_dyn` Stata and R commands can be used. The `did_had` package may only be helpful to estimate the treatment effects of the last treatment cohort.

6 Applications

6.1 Effect on US employment of eliminating potential tariffs spikes on imports from China.

Research question. In 2001, the United States (US) granted Permanent Normal Trade Relations (PNTR) to China, thus eliminating the possibility of tariff spikes: prior to 2001, congress had to re-approve every year the low tariff rates Chinese imports were subjected to. The treatment in this application is the magnitude of the potential tariffs' spike eliminated by the reform, namely the difference between the non-NTR and NTR tariff rates. This treatment varies substantially across industries, taking values ranging from 2 to 64 percentage points, with a mean and standard deviation respectively equal to 30 and 14 percentage points.

Scope of our re-analysis. The data used by Pierce and Schott (2016) is proprietary, except that used to produce their Table 3. We focus on Column (3) therein: Column (2) is a placebo looking at the effect of the treatment on employment in Europe, while Column (1) is a triple difference comparing the effect of the treatment in the US and in Europe.

TWFE regressions. In their Table 3 Column (3), the authors use a panel of 103 US industries from 1997 to 2002 and from 2004 to 2005,[5] and regress the log employment of industry g in year t $Y_{g,t}$ on industry and year fixed effects and on $D_{g,t}$, weighting the regression by industries' 1997 employment. They find a negative and significant coefficient (-0.65, s.e.=0.27), suggesting that eliminating a potential tariff spike of 100 percentage points reduces US employment by 0.65 percentage point. To fit in the setting considered in this paper, we consider unweighted regressions of $Y_{g,t} - Y_{g,2000}$, for $t \in \{2001, 2002, 2004, 2005\}$, on $D_{g,2001}$. As $G = 103$ is not large, we follow the recommendations from Imbens and Kolesar (2016) and use HC2 standard errors with the DOF adjustment recommended by Bell and McCaffrey (2002) to obtain more reliable confidence intervals. Panel A of Table

[5] As noted by Pierce and Schott (2016), 2003 data is missing for all US industries in the UNIDO dataset used to produce the table. While the version of the UNIDO dataset downloaded by the authors had 104 US industries, the version we downloaded in 2023 has 103 industries, presumably due to some industry regrouping.

2 below shows that $\widehat{\beta}_{fe,t}$ is small and insignificant in t =2001, before becoming large and significant in t =2002, and even larger in t =2004 and t =2005. The TWFE coefficients in Table 2 are all less negative than that in Table 3 Column (3) of the paper. This is because our regressions are not weighted. With weighting, some of our coefficients become more negative than the coefficient in the paper.

Placebo tests of Assumption 2 are rejected. To assess the reliability of the TWFE regressions in Panel A of Table 2, we begin by placebo-testing Assumption 2. First, Panel B of Table 2 shows regressions of $Y_{g,t} - Y_{g,2000}$, for $t \in \{1999, 1998, 1997\}$, on $D_{g,2001}$. All those placebo estimators are statistically significant, though their magnitude is smaller than that of the actual estimators in Panel A. This suggests that Assumption 2 might be violated, though it does not seem that violations of Assumption 2 can fully account for the estimated effects.[6] Panel B of Table 2 also shows p-values of Stute tests that $E(Y_{g,t} - Y_{g,2000}|D_{g,2001}) = \mu_t$, for $t \in \{1999, 1998, 1997\}$. The null is rejected at the 5% level in 1998, and the joint test is rejected at the 10% level (p-value=0.06).

Pre-trends tests of Assumption 2 are no longer rejected when industry-specific linear trends are controlled for. The pre-trends estimators increase as we look at employment evolutions over a longer horizon. Then, the violation of Assumption 2 in this

[6] Those findings are at odds with those from Figure 2 in Pierce and Schott (2016). Therein, the authors run the same pre-trends tests as in Panel B of Table 2, on the proprietary dataset they use for most of their analysis, where industries are defined at a more disaggregated level than in our data, and they do not find statistically significant pre-trends. Their regressions are weighted unlike ours, but if we weight our regressions by industries' 1997 employment, pre-trends tests are still rejected. It seems that while disaggregated treatments are uncorrelated with industries' employment pre-trends, the aggregated variables are correlated, a version of the so-called "ecological inference problem". By the law of total covariance, for any random variables U, V, and W,

$$\text{Cov}(U, V) = \text{Cov}(E(U|W), E(V|W)) + E(\text{Cov}(U, V|W)).$$

Letting U and V respectively denote disaggregated treatments and employment pre-trends, and W be a variable aggregating industries when going from the proprietary to the publicly available data, the insignificant pre-trends in the paper ($\text{Cov}(U, V) = 0$) could be due to negative pre-trends across aggregated industries ($\text{Cov}(E(U|W), E(V|W)) < 0$), compensated by positive pre-trends within industries ($E(\text{Cov}(U, V|W)) > 0$). Those offsetting differential trends are not a concern for the results in Table 1 of Pierce and Schott (2016), which require that Assumption 2 hold at the aggregation level in the table.

data might be due to industry-specific linear trends that are correlated with industries' treatments, and Assumption 2 might be plausible when such linear trends are controlled for. Accordingly, we replace Assumption 2 by

$$E(Y_{g,t}(0) - Y_{g,2000}(0) - (t - 2000) \times (Y_{g,2000}(0) - Y_{g,1999}(0))|D_{g,2001}) = \mu_t. \qquad (18)$$

$Y_{g,2000}(0) - Y_{g,1999}(0)$ captures industry g's linear trend without treatment. Then, $Y_{g,t}(0) - Y_{g,2000}(0) - (t - 2000) \times (Y_{g,2000}(0) - Y_{g,1999}(0))$ is g's deviation from its linear trend from 2000 to t. Therefore, (18) requires that industries' deviations from their linear trend are mean-independent from the NTR-gap treatment. Under this assumption, treatment-effect estimators can be obtained by regressing, for $t \in \{2001, 2002, 2004, 2005\}$, $Y_{g,t} - Y_{g,2000} - (t-2000) \times (Y_{g,2000} - Y_{g,1999})$ on $D_{g,2001}$. Similarly, to placebo-test (18), one can regress, for $t \in \{1998, 1997\}$, $Y_{g,t} - Y_{g,1999} - (t-1999) \times (Y_{g,2000} - Y_{g,1999})$ on $D_{g,2001}$. Panel C of Table 2 below shows that this placebo test is not rejected. That panel also shows p-values of Stute tests that $E(Y_{g,t} - Y_{g,1999} - (t-1999) \times (Y_{g,2000} - Y_{g,1999})|D_{g,2001}) = \mu_t$, for $t \in \{1998, 1997\}$. Those tests are not rejected, and the joint test is also not rejected. This suggests that Assumption 2 holds when industry-specific linear trends are accounted for.

TWFE estimators with industry-specific linear trends are still negative but smaller than without linear trends, and less significant. TWFE estimators with industry-specific linear trends are shown in Panel D of Table 2. They are smaller and less significant than estimators without linear trends, but the estimated effect in 2004 is still significant at the 5% level, and that in 2002 is significant at the 10% level.

Under (18) alone, TWFE regressions do not estimate convex combinations of effects. We follow (4),[7] and estimate the weights attached to the TWFE coefficients in Panel D of Table 2. Estimation is carried out with the `twowayfeweights` Stata package (de Chaisemartin et al., 2019). We find that those coefficients estimate a weighted sum of the effects of the PNTR in the 103 industries, where 62 estimated weights are strictly positive, while 41 are strictly negative. The negative weights sum to -0.32.

Tests of the homogeneous and linear effect assumption are not rejected. To test if heterogeneous effects could bias the TWFE regressions with linear trends, we

[7] This decomposition applies to the regression of $Y_{g,t} - Y_{g,1999} - (t-1999) \times (Y_{g,2000} - Y_{g,1999})$ on $D_{g,2}$.

run the Stute test of linearity, on $Y_{g,t} - Y_{g,2000} - (t - 2000) \times (Y_{g,2000} - Y_{g,1999})$, for $t \in \{2001, 2002, 2004, 2005\}$. Panel D of Table 2 below shows that of those 4 tests, one is rejected at the 10% level, but the joint test is not rejected (p-value=0.40). As the interpretation of this test depends on whether there are quasi-stayers in this application, we conduct the test we propose in Section 3.4. We find that $D^2_{2,(1)}/(D^2_{2,(2)} - D^2_{2,(1)}) = 2.84$, so the test is far from being rejected at all conventional levels (p-value=0.26). $D_{2,(1)}/(D_{2,(2)} - D_{2,(1)}) = 6.15$, so a less conservative and more powerful test that assumes a strictly positive density is also not rejected (p-value=0.14). Thus it seems that there are quasi-stayers in this application, in which case there is an "if and only if" between the null of the Stute test and the homogeneous and linear effect assumption.

Heterogeneity-robust estimators. While the joint Stute test is not rejected in Panel D of Table 2, with 103 observations and four years of data the test may lack power. The test would be more powerful if we could run it on the authors' proprietary dataset, which has 315 industries and many more years of data. Unfortunately, this dataset is built from the Longitudinal Business Database, which can only be accessed by US citizens or permanent residents from a US Federal Statistical Research Data Center. Instead, and as a further robustness check, we also compute an heterogeneity-robust estimator. As this application has quasi-stayers, in principle we could compute our heterogeneity-robust estimator with quasi-stayers $\widehat{\beta}^{qs}_{\widehat{h}^*_G}$. However, with a sample size of 103, this non-parametric estimator is very imprecise. Instead, we compute our parametric estimator $\widehat{\beta}^{ns}$, accounting for industry-specific linear trends, and assuming that $E[\text{TE}_2|D_2 = d] = \delta_0 + \delta_1 d$, thus allowing for some CAS heterogeneity. Standard errors are computed using the bootstrap. Estimated effects are larger than using the TWFE regression, especially in 2002. The effect that year is now significant at all conventional levels. On the other hand, the effect in 2004 is not significant anymore. By and large, these results confirm those based on TWFE regressions.

Table 2: Effects of Eliminating Potential Tariff Spikes on Chinese Imports

	Panel A: Effects			
	2001	2002	2004	2005
$\widehat{\beta}_{fe}$	-0.06	-0.26	-0.54	-0.53
95% CI	$[-0.14, 0.02]$	$[-0.41, -0.11]$	$[-0.85, -0.23]$	$[-0.87, -0.19]$
	Panel B: Placebo effects			
	1999	1998	1997	
$\widehat{\beta}_{fe}$	0.06	0.14	0.16	
95% CI	$[0.00, 0.11]$	$[0.03, 0.25]$	$[0.02, 0.31]$	
P-value Stute test of mean indep.	0.15	0.02	0.15	
P-value joint test of mean indep.		0.06		
	Panel C: Placebo effects with industry-specific linear trends			
	1998	1997		
$\widehat{\beta}_{fe}$	-0.02	-0.05		
95% CI	$[-0.093, 0.044]$	$[-0.136, 0.044]$		
P-value Stute test of mean indep.	0.30	0.51		
P-value joint test of mean indep.	0.47			
	Panel D: Effects with industry-specific linear trends			
	2001	2002	2004	2005
$\widehat{\beta}_{fe}$	-0.00	-0.14	-0.31	-0.24
95% CI	$[-0.08, 0.08]$	$[-0.30, 0.01]$	$[-0.62, 0.00]$	$[-0.58, 0.10]$
P-value Stute test of linearity	0.38	0.06	0.38	0.53
P-value joint test of linearity		0.40		
$\widehat{\beta}^{ns}$	-0.02	-0.38	-0.43	-0.38
95% CI	$[-0.17, 0.14]$	$[-0.66, -0.11]$	$[-1.15, 0.29]$	$[-1.11, 0.34]$
Observations	103	103	103	103

Notes: This table shows estimated effects (Panels A and D) and placebo effects (Panels B and C) on US employment of eliminating potential tariffs spikes on imports from China. Estimation uses log employment data for a panel of 103 US industries from 1997 to 2002 and from 2004 to 2005. In Panels A and B, TWFE regressions are shown. In Panels C and D, TWFE regressions with industry-specific linear trends are shown. In Panels B, C, D, we also show p-values of Stute tests of mean independence and linearity. Finally, in Panel D we also show an heterogeneity-robust estimator, relying on a parametric functional form assumption.

6.2 Effect of the exposure rate to independent information on voting behavior

Research question. In 1996, a new TV channel called NTV was introduced in Russia. At that time, it was the only TV channel not controlled by the government. Enikolopov et al. (2011) study the effect of having access to this independent news source on voting behavior, using voting outcomes for 1,938 Russian subregions in the 1995 and 1999 election, and the fact that after 1996, NTV's exposure rate is heterogeneous across regions: this rate ranges from 28 to 91%, with an average of 58% and a standard deviation of 9%.

Scope of our re-analysis and TWFE regressions. Our re-analysis is concerned with Table 3 in Enikolopov et al. (2011), where the authors use $\widehat{\beta}_{fe}$ to estimate the NTV-exposure-rate's effect on five outcomes: the share of the electorate voting for the SPS and Yabloko parties, two opposition parties supported by NTV; the share of the electorate voting for the KPRF and LDPR parties, two parties not supported by NTV; and electoral turnout. They find that $\widehat{\beta}_{fe} = 6.65$ (s.e.$= 1.40$) for the SPS voting rate, and $\widehat{\beta}_{fe} = 1.84$ (s.e.$= 0.76$) for the Yabloko voting rate. According to these regressions, increasing the NTV exposure rate from 0 to 100% increases the share of votes for the SPS and Yabloko opposition parties by 6.65 and 1.84 percentage points, respectively. $\widehat{\beta}_{fe}$ is small and insignificant for the remaining three outcomes.

Under Assumption 2 alone, $\widehat{\beta}_{fe}$ does not estimate a convex combination of effects. We follow (4), and estimate the weights attached to the coefficients $\widehat{\beta}_{fe}$, without making further assumptions than Assumption 2. Estimation is carried out with the `twowayfeweights` Stata package. We find that $\widehat{\beta}_{fe}$ estimates a weighted sum of the effects of NTV in 1999 in the 1,938 regions, where 918 estimated weights are strictly positive, while 1,020 are strictly negative. The negative weights sum to -2.26. The same decomposition of $\widehat{\beta}_{fe}$ in this application was already reported in Section 2.2 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2020).

The test of Assumption 2 and (5) is rejected for four outcomes out of five. In spite of its negative weights, $\widehat{\beta}_{fe}$ is still consistent for AS if on top of Assumption 2, (5) also holds. The Stute test is not rejected for the KPRF outcome, but it is rejected at the 5% level for all other outcomes, thus strongly suggesting that Assumption 2 and (5) do not hold. Unfortunately, the data does not contain electoral outcomes for another election before the introduction of NTV, so we cannot run a pre-trends test of Assumption 2.

Heterogeneity-robust estimators. Our first set of results suggests that $\widehat{\beta}_{fe}$ may not be a reliable estimator in this application. Then, one may want to use another estimator. Which estimator we should use depends on whether there are quasi-stayers in this application, so we conduct the test we propose in Section 3.4. We find that $D^2_{2,(1)}/(D^2_{2,(2)} - D^2_{2,(1)}) = 25.51$ and $D_{2,(1)}/(D_{2,(2)} - D_{2,(1)}) = 51.51$, so our two tests are rejected at the 5% level (p-value=0.04 and 0.02, respectively). As this suggests that there are no quasi-stayers, we first bound WAS under a boundedness condition on the size of the treatment effect, before imposing parametric assumptions to recover point identification.

The non-parametric bounds for WAS in Theorem 4 are uninformative. We use the `clr` Stata package (Chernozhukov et al., 2015) to compute half-median-unbiased estimators of $\sup_{d_2 \in \text{Supp}(D_2)} \theta_\ell(d_2)$ and $\inf_{d_2 \in \text{Supp}(D_2)} \theta_u(d_2)$ in (15), following Chernozhukov et al. (2013). Then, we replace expectations by sample averages for all the other quantities in (15), and finally obtain estimators $\widehat{B}_-$ and $\widehat{B}_+$ of the bounds in Theorem 4. For the SPS voting rate, we let $\underline{M} = -13$ and $\overline{M} = 13$. $\widehat{\beta}_{fe} = 6.65$ for this outcome, so 6.65 is presumably a treatment-effect magnitude that was deemed plausible by the paper's authors and readers. With $\underline{M} = -13$ and $\overline{M} = 13$, Assumption 4 means that the CAS's absolute values are no larger than about twice this value. This still allows for substantial treatment effect heterogeneity, while also restricting substantially the CAS's range of possible values: without imposing any restriction, that range is $[-100, 100]$. We find that $\widehat{B}_- = -2.88$ and $\widehat{B}_+ = 9.99$: even without taking into account the bounds' estimation error, we cannot reject WAS = 0. The largest bound on CAS absolute values such that $0 \notin [\widehat{B}_-, \widehat{B}_+]$ is 6.97, a value fairly close to $\widehat{\beta}_{fe}$. We apply the same methodology to bound WAS for the Yabloko voting rate, the other outcome for which the authors find a significant TWFE coefficient, letting $\overline{M} = 4 \approx 2\widehat{\beta}_{fe}$ and $\underline{M} = -4 \approx -2\widehat{\beta}_{fe}$. There as well, $0 \in [\widehat{B}_-, \widehat{B}_+]$.

Estimators more robust to heterogeneous effects are different from $\widehat{\beta}_{fe}$. In Table 3 below, we compute $\widehat{\beta}^{ns}$ assuming that $E[\text{TE}_2|D_2 = d] = \delta_0 + \delta_1 d$, thus allowing for some CAS heterogeneity. $\widehat{\beta}^{ns}$ is close to $\widehat{\beta}_{fe}$ but not significantly different from zero for the SPS vote outcome. $\widehat{\beta}^{ns}$ is seven times larger than $\widehat{\beta}_{fe}$ and significantly different from zero for the Yabloko vote outcome. $\widehat{\beta}^{ns}$ and $\widehat{\beta}_{fe}$ are both insignificantly different from zero for the KPRF vote outcome. $\widehat{\beta}^{ns}$ is large, negative, and significant for the LDPR vote and Turnout outcomes, while $\widehat{\beta}_{fe}$ is insignificant for those outcomes. For all outcomes, $\widehat{\beta}^{ns}$ is much more noisy than $\widehat{\beta}_{fe}$. Thus, allowing for heterogeneous effects, even in a fairly restricted way, yields noisy estimates, that often differ from the authors' original estimates. Note that for the LDPR vote and Turnout outcomes, $\widehat{\beta}^{ns}$ is implausibly large, which may indicate a violation of the assumptions underlying that estimator.

Table 3: Effects of Exposure Rate to Independent Information on Voting

	SPS vote	Yabloko vote	KPRF vote	LDPR vote	Turnout
$\widehat{\beta}_{fe}$	6.65	1.84	-2.20	1.18	-2.06
(s.e.)	(1.40)	(0.76)	(2.12)	(1.38)	(2.01)
P-value Stute test	0.026	0.016	0.138	$< 10^{-3}$	0.002
$\widehat{\beta}^{ns}$	5.11	12.78	12.80	-39.18	-28.19
(s.e.)	(6.46)	(4.51)	(10.36)	(7.04)	(9.27)
Observations	1,938	1,938	1,938	1,938	1,938

Notes: This table shows estimated effects of the exposure rate to independent information on voting outcomes in Russia, using voting data for 1,938 Russian subregions in the 1995 and 1999 elections. In the first line, effects are estimated using TWFE regressions, while in the third line they are estimated using an heterogeneity-robust estimator, relying on a parametric functional form assumption.

7 Conclusion

We consider treatment-effect estimation in heterogeneous adoption designs without stayers, where no group is treated at period one, and groups receive a strictly positive treatment dose at period two. Under a parallel-trends assumption, the commonly-used two-way fixed effects estimator may not be robust to heterogeneous treatment effects in those designs. We start by providing a test of a condition under which this estimator is robust. When this test is rejected, we propose alternative estimators that are robust to heterogeneous treatment effects. We use our results to revisit Pierce and Schott (2016) and Enikolopov et al. (2011). While in Pierce and Schott (2016), the findings from TWFE regressions seem robust, this does not appear to be the case in Enikolopov et al. (2011).

References

Bell, R. M. and D. F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*(2), 169–182.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2021). Difference-in-differences with a continuous treatment.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association 113*(522), 767–779.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. arXiv preprint arXiv:1906.00198.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Chernozhukov, V., W. Kim, S. Lee, and A. M. Rosen (2015). Implementing intersection bounds in stata. *The Stata Journal 15*(1), 21–44.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: Estimation and inference. *Econometrica 81*(2), 667–737.

de Chaisemartin, C., D. Ciccia, X. D'Haultfoeuille, F. Knau, and D. Sow (2024a). Didhad: R module to estimate the effect of a treatment on an outcome in a heterogeneous adoption design with no stayers but some quasi stayers.

de Chaisemartin, C., D. Ciccia, X. D'Haultfoeuille, F. Knau, and D. Sow (2024b). Didhad: Stata module to estimate the effect of a treatment on an outcome in a heterogeneous adoption design with no stayers but some quasi stayers.

de Chaisemartin, C., D. Ciccia, X. D'Haultfoeuille, F. Knau, and D. Sow (2024c). Stutetest: R module to perform stute (1997) linearity test.

de Chaisemartin, C., D. Ciccia, X. D'Haultfoeuille, F. Knau, and D. Sow (2024d). Stute_test: Stata module to perform stute (1997) linearity test.

de Chaisemartin, C. and X. D'Haultfœuille (2015). Fuzzy differences-in-differences. ArXiv e-prints, eprint 1510.01757v2.

de Chaisemartin, C. and X. D'Haultfœuille (2018). Fuzzy differences-in-differences. *The Review of Economic Studies 85*(2), 999–1028.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–2996.

de Chaisemartin, C. and X. D'Haultfœuille (2021). Difference-in-differences estimators of intertemporal treatment effects. arXiv preprint arXiv:2007.04267.

de Chaisemartin, C., X. D'Haultfoeuille, and A. Deeb (2019). Twowayfeweights: Stata module to estimate the weights and measure of robustness to treatment effect heterogeneity attached to two-way fixed effects regressions. Technical report, HAL.

de Chaisemartin, C., X. D'Haultfœuille, and G. Vazquez-Bare (2024). Difference-in-difference estimators with continuous treatments and no stayers. In *AEA Papers and Proceedings*, Volume 114, pp. 610–613. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

de Chaisemartin, C. and X. D'Haultfoeuille (2024). Under the null of valid specification, pre-tests of valid specification do not distort inference. arXiv preprint arXiv:2407.03725.

de Chaisemartin, C., X. D'Haultfoeuille, F. Pasquier, and G. Vazquez-Bare (2022). Difference-in-differences estimators of the effect of a continuous treatment. arXiv preprint arXiv:2201.06898.

Duggan, M. and F. Scott Morton (2010). The effect of medicare part d on pharmaceutical prices and utilization. *American Economic Review 100*(1), 590–607.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and political persuasion: Evidence from russia. *American Economic Review 101*(7), 3253–3285.

Fetzer, T. (2019). Did austerity cause brexit? *American Economic Review 109*(11), 3849–3886.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies 79*(3), 933–959.

Imbens, G. and Y. Xu (2024). Lalonde (1986) after nearly four decades: Lessons learned. arXiv preprint arXiv:2406.00827.

Imbens, G. W. and M. Kolesar (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics 98*(4), 701–712.

Imbens, G. W., D. B. Rubin, and B. I. Sacerdote (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American economic review 91*(4), 778–794.

Pierce, J. R. and P. K. Schott (2016). The surprisingly swift decline of us manufacturing employment. *American Economic Review 106*(7), 1632–62.

Sasaki, Y. and T. Ura (2021). Slow movers in panel data. arXiv preprint arXiv:2110.12041.

Shorack, G. and J. Wellner (1986). *Empirical Processes with Applications to Statistics.* John Wiley and Sons.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics 25*(2), 613–641.

Stute, W., W. G. Manteiga, and M. P. Quindimil (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association 93*(441), 141–149.

van der Vaart, A. (2000). *Asymptotics Statistics.* Cambridge University Press.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics letters 57*(2), 135–143.

A Proofs

A.1 Theorem 2

1. Asymptotic size control.

Fix $k \in \{1, 2\}$, a sequence $(x_G)_{G\geq 1}$ with $x_G > 0$ and $x_G \to 0$; without loss of generality, we assume hereafter that $x_G \leq \overline{d}$ for all G. Fix also a sequence $(F_G)_{G\geq 1}$, with $F_G \in \mathcal{F}^{k,0,\overline{d}}_{m,M}$ for all $G \geq 1$. Consider any subsequence $(F_{G'})_{G'\geq 1}$ of $(F_G)_{G\geq 1}$. Since $F^{(k)}_G$ is Lipschitz with parameter M on $[0, \overline{d}]$, the subsequence $(F^{(k)}_{G'})_{G'\geq 1}$ is equicontinuous. Then, by Arzelà–Ascoli theorem, it admits a further subsequence that converges uniformly; let $H_k(\cdot)$ denote its limit. We now separate the cases $k = 1$ and $k = 2$.

Case $k = 1$. By the mean value theorem,

$$F_{G''}(F^{-1}_{G''}(x_{G''})) = F'_{G''}(d_{G''})F^{-1}_{G''}(x_{G''}),$$

for some $d_{G''} \in [0, F^{-1}_{G''}(x_{G''})]$. Hence, $F'_{G''}(d_{G''}) \neq 0$ and

$$\frac{F^{-1}_{G''}(x_{G''})}{x_{G''}} = \frac{1}{F'_{G''}(d_{G''})}.$$

Because $F_G(d) \geq md$, $F^{-1}_G(u) \leq u/m$. Thus, $F^{-1}_{G''}(x_{G''}) \to 0$, which implies that $d_{G''} \to 0$. Then, by uniform convergence of $F'_{G''}$ and $H_1(0) \geq m$ (in view of $F'_G(0) \geq m$), we obtain

$$\frac{F^{-1}_{G''}(x_{G''})}{x_{G''}} \to \frac{1}{H_1(0)}.$$

For U uniformly distributed over $[0, 1]$, $F^{-1}_G(U) \overset{d}{=} D_2$. As a result, by the representation of spacings (see, e.g., Shorack and Wellner, 1986, p.721),

$$\left(D_{2,(1)}, D_{2,(2)} - D_{2,(1)}\right) \overset{d}{=} \left(F^{-1}_G\left(\frac{E_1}{\sum_{i=1}^{G+1} E_i}\right), F^{-1}_G\left(\frac{E_1 + E_2}{\sum_{i=1}^{G+1} E_i}\right) - F^{-1}_G\left(\frac{E_1}{\sum_{i=1}^{G+1} E_i}\right)\right), \quad (19)$$

where $(E_1, ..., E_{G+1})$ are i.i.d. and follow an Exponential(1) distribution. Also, by the law of large numbers,

$$\frac{E_1}{\sum_{i=1}^{G+1} E_i} \xrightarrow{p} 0, \quad \frac{E_1 + E_2}{\sum_{i=1}^{G+1} E_i} \xrightarrow{p} 0.$$

Hence, by the extended continuous mapping theorem (see, e.g., van der Vaart, 2000, Theorem 18.11),

$$\begin{aligned}
&G''\left(F_{G''}^{-1}\left(\frac{E_1}{\sum_{i=1}^{G''+1}E_i}\right),\ F_{G''}^{-1}\left(\frac{E_1+E_2}{\sum_{i=1}^{G''+1}E_i}\right)-F_{G''}^{-1}\left(\frac{E_1}{\sum_{i=1}^{G''+1}E_i}\right)\right)\\
=&\left(\frac{E_1}{\frac{1}{G''}\sum_{i=1}^{G''+1}E_i}\times\frac{F_{G''}^{-1}\left(\frac{E_1}{\sum_{i=1}^{G''+1}E_i}\right)}{E_1/\sum_{i=1}^{G''+1}E_i},\frac{E_1+E_2}{\frac{1}{G''}\sum_{i=1}^{G''+1}E_i}\times\frac{F_{G''}^{-1}\left(\frac{E_1+E_2}{\sum_{i=1}^{G''+1}E_i}\right)}{(E_1+E_2)/\sum_{i=1}^{G''+1}E_i}\right.\\
&\left.-\frac{E_1}{\frac{1}{G''}\sum_{i=1}^{G''+1}E_i}\times\frac{F_{G''}^{-1}\left(\frac{E_1}{\sum_{i=1}^{G''+1}E_i}\right)}{E_1/\sum_{i=1}^{G''+1}E_i}\right)\\
\xrightarrow{p}&\frac{1}{H_1(0)}\times(E_1,E_2).
\end{aligned}$$

Combined with (19) and, again, the continuous mappping theorem, we obtain

$$\frac{D_{2,(1)}}{D_{2,(2)}-D_{2,(1)}}\xrightarrow{d}\frac{E_1}{E_2}, \tag{20}$$

where the convergence should be understood along the previous subsequence. Then, using $D_{2,(1)}/[D_{2,(2)}+D_{2,(1)}]\leq 1$, we obtain

$$\begin{aligned}
P_{F_{G''}}(W_\alpha)&=P_{F_{G''}}(T>1/\alpha-1)\\
&\leq P_{F_{G''}}\left(\frac{D_{2,(1)}}{D_{2,(2)}-D_{2,(1)}}>1/\alpha-1\right)\\
&\to\alpha.
\end{aligned}$$

Now, let $u_G:=\max(0,P_{F_G}(W_\alpha)-\alpha)$. The previous display proves that $u_{G''}\to 0$. By Urysohn subsequence principle, this implies that $u_G\to 0$. Since the sequence $(F_G)_{G\geq 1}$ was arbitrary, $\limsup_{G\to\infty}\sup_{F\in\mathcal{F}^{1,0,\bar{d}}_{m,M}}P_F(W_\alpha)\leq\alpha$.

Case $k=2$. In this case, $F'_{G''}(0)=0$. Then, by the mean value theorem again,

$$F_{G''}(F_{G''}^{-1}(x_{G''}))=\frac{F''_{G''}(d_{G''})}{2}F_{G''}^{-1}(x_{G''})^2, \tag{21}$$

for some $d_{G''}\in[0,F_{G''}^{-1}(x_{G''})]$. This implies that $F''_{G''}(d_{G''})>0$ and

$$\frac{F_{G''}^{-1}(x_{G''})^2}{x_{G''}}=\frac{2}{F''_{G''}(d_{G''})}.$$

Because $F_G(d) \geq md^2/2$, $F_G^{-1}(u) \leq (2u/m)^{1/2}$. Thus, $F_{G''}^{-1}(x_{G''}) \to 0$ and so $d_{G''} \to 0$. By uniform convergence of $F''_{G''}$ and $H_2(0) \geq m$ (since $F''_G(0) \geq m$), we obtain

$$\frac{F_{G''}^{-1}(x_{G''})^2}{x_{G''}} \to \frac{2}{H_2(0)}. \tag{22}$$

Let $(E_1, ..., E_{G+1})$ be as above. As in (19),

$$\left(D^2_{2,(1)}, D^2_{2,(2)} - D^2_{2,(1)}\right) \overset{d}{=} \left(F_G^{-1}\left(\frac{E_1}{\sum_{i=1}^{G+1} E_i}\right)^2, F_G^{-1}\left(\frac{E_1+E_2}{\sum_{i=1}^{G+1} E_i}\right)^2 - F_G^{-1}\left(\frac{E_1}{\sum_{i=1}^{G+1} E_i}\right)^2\right).$$

Then, by the same reasoning as to obtain (20), but using (22) and the previous display, we get

$$\frac{D^2_{2,(1)}}{D^2_{2,(2)} - D^2_{2,(1)}} \xrightarrow{d} \frac{E_1}{E_2}.$$

As a result, $P_{F_{G''}}(W_\alpha) \to \alpha$. By Urysohn subsequence principle, this implies that $P_{F_G}(W_\alpha) \to \alpha$. Since the sequence $(F_G)_{G\geq 1}$ was arbitrary, we obtain $\limsup_{G\to\infty} \sup_{F \in \mathcal{F}^{2,0,\overline{d}}_{m,M}} P_F(W_\alpha) = \alpha$.

2. Uniform consistency

If $F \in \mathcal{F}^{k,\underline{d},\overline{d}}_{m,M}$ for some $k \in \{1, 2\}$ and $\underline{d} > 0$, we have

$$T = \frac{(\underline{d} + \tilde{D}_{2,(1)})^2}{(\tilde{D}_{2,(2)} - \tilde{D}_{2,(1)}(2\underline{d} + \tilde{D}_{2,(2)} + \tilde{D}_{2,(1)})}, \tag{23}$$

where the cdf of $\tilde{D}_{2,k}$ belongs to $\mathcal{F}^{k,0,\overline{d}-\underline{d}}_{m,M}$. We reason as above by considering sequences F_G in $\mathcal{F}^{k,0,\overline{d}-\underline{d}}_{m,M}$ and appropriate subsequences for which $F_G^{(k)}$ converges uniformly. By Point 1 above, $(\tilde{D}_{2,(1)}, \tilde{D}_{2,(2)}) \xrightarrow{p} 0$, along such subsequences. By (23), this implies that, still along such subsequences, $T \xrightarrow{p} \infty$. Point 2 follows by Urysohn subsequence principle again.

3. Local power with positive density at the boundary

As above, we consider a sequence F_G of cdfs and we prove that $u_G := \min(0, P_{F_G}(W_\alpha) - \alpha) \to 0$. For any subsequence, we consider a further subsequence (G'') such that (i) the kth derivative of the cdf of $\tilde{D}_{2,k}$ converges and (ii) the sequence $(G''\underline{d}_{G''})$ converges; we denote by $\lambda \in [1, \infty]$ its limit. Then, using (23) but with $\underline{d}$ replaced by $\underline{d}_G$ and the same reasoning as to get (20), we obtain, after some algebra and along the subsequence (G''),

$$T \xrightarrow{d} T_\lambda := \frac{(\lambda + E_1)^2}{E_2(2\lambda + E_2 + 2E_1)},$$

with the understanding that $T_\infty = \infty$. In this latter case, the same reasoning as in Point 2 above shows that $P_{F_{G''}}(W_\alpha) \to 1$ and thus $u_{G''} \to 0$. Now, assume that $\lambda < \infty$. Remark that $T_\lambda > 1/\alpha - 1$ if and only if $E_1 < r_1(E_2)$ or $E_1 > r_2(E_2)$, where

$$\begin{aligned} r_1(x) &= -\lambda + x(s - \sqrt{s(s+1)}), \\ r_2(x) &= -\lambda + x(s + \sqrt{s(s+1)}), \end{aligned}$$

with $s = 1/\alpha - 1$. Actually, since $r_1(x) < 0$ for all $x \geq 0$, $T_\lambda > 1/\alpha - 1$ if and only if $E_1 > r_2(E_2)$. Then, we obtain after some algebra that

$$P(T_\lambda > s) = 1 + \left(\frac{1}{\gamma} - 1\right) \exp(-\lambda/(\gamma - 1)),$$

where $\gamma = s + 1 + \sqrt{s(s+1)} = 1/\alpha + \sqrt{1/\alpha(1/\alpha - 1)}$. The right-hand side is increasing in λ. Moreover, some elementary but tedious algebra shows that for $\lambda = 1$, $\alpha \mapsto P(T_1 > 1/\alpha - 1) - \alpha$ is strictly increasing and thus (since $\lim_{\alpha \to 0} P(T_1 > 1/\alpha - 1) = 0$), $P(T_1 > 1/\alpha - 1) > \alpha$ for all $\alpha > 0$. Then, for all $\lambda \geq 1$, $P(T_\lambda > 1/\alpha - 1) \geq P(T_1 > 1/\alpha - 1) > \alpha$. Let $u_G := \min(0, P_{F_G}(W_\alpha) - \alpha)$. The previous display proves that $u_{G''} \to 0$. By Urysohn subsequence principle, this implies that $u_G \to 0$. The result follows.

4. Local power with null density at the boundary

We reason as in Point 3 so we just highlight the differences here. Assuming that $G''^{1/2}\underline{d}_{G''} \to \lambda > 0$,

$$T \xrightarrow{d} T'_\lambda := \frac{\left(\lambda + E_1^{1/2}\right)^2}{\left(\lambda + (E_1 + E_2)^{1/2}\right)^2 - \left(\lambda + E_1^{1/2}\right)^2},$$

again with the understanding that $T'_\infty = \infty$. The derivative of $\lambda \mapsto T'_\lambda$ is strictly positive, so $T'_\lambda > T'_0 = E_1/E_2$. As a result,

$$P(T'_\lambda > s) > P(E_1/E_2 > s) = \alpha.$$

The result follows with the same reasoning as above.

A.2 Theorem 4.

First, we have

$$\text{WAS} = \frac{E\left[Y_2 - Y_1 - \mu_0\right]}{E\left[D_2\right]}.$$

Moreover, for all $d_2 \in \text{Supp}(D_2)$,

$$\mu_0 = m(d_2) - d_2 E[\text{TE}_2|D_2 = d_2]. \tag{24}$$

In view of (24), the constraint $\underline{M} \leq E[\text{TE}_2|D_2] \leq \overline{M}$ is equivalent to

$$\forall d_2 \in \text{Supp}(D_2),\ \underline{M} \leq \frac{m(d_2) - \mu_0}{d_2} \leq \overline{M},$$

which is equivalent to

$$\forall d_2 \in \text{Supp}(D_2),\ m(d_2) - \overline{M} d_2 \leq \mu_0 \leq m(d_2) - \underline{M} d_2.$$

Therefore, $\mu_0 \in [\underline{\mu}, \overline{\mu}]$. This proves the validity of the bounds, and also proves that Assumptions 2 and 4 cannot jointly hold if $\overline{\mu} < \underline{\mu}$.

We now prove that the lower bound is sharp, by exhibiting a DGP compatible with the data and satisfying Assumptions 2 and 4, for which WAS $= B_-$. $\underline{E}$ denotes expectations under this DGP. Let $\underline{E}[\text{TE}_2|D_2 = d_2] = \frac{m(d_2)-\underline{\mu}}{d_2}$ and $\underline{E}[\Delta Y(0)|D_2 = d_2] = \underline{\mu}$. This DGP satisfies Assumption 2. It follows from the definition of $\overline{\mu}$ that $\underline{M} \leq \underline{E}[\text{TE}_2|D_2]$, and it follows from $\underline{\mu} \leq \overline{\mu}$ and the definition of $\underline{\mu}$ that $\underline{E}[\text{TE}_2|D_2] \leq \overline{M}$, so this DGP also satisfies Assumption 4.

$$\underline{E}[\Delta Y|D_2 = d_2] = \underline{\mu} + d_2 \underline{E}[\text{TE}_2|D_2 = d_2] = m(d_2),$$

so this DGP is compatible with the data. Finally,

$$E\left[\frac{D_2}{E[D_2]} \underline{E}[\text{TE}_2|D_2]\right] = B_-.$$

One could follow similar steps to prove that the upper bound is sharp.

CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

The Center for Research in Economics and Statistics (CREST) is a leading French scientific institution for advanced research on quantitative methods applied to the social sciences.

CREST is a joint interdisciplinary unit of research and faculty members of CNRS, ENSAE Paris, ENSAI and the Economics Department of Ecole Polytechnique. Its activities are located physically in the ENSAE Paris building on the Palaiseau campus of Institut Polytechnique de Paris and secondarily on the Ker-Lann campus of ENSAI Rennes.