Testing for equivalence of pre-trends in Difference-in-Differences estimation

Holger Dette
Department of Mathematics, Ruhr University Bochum
and
Martin Schumann
School of Business and Economics, Maastricht University

February 17, 2023

Abstract

The plausibility of the "parallel trends assumption" in Difference-in-Differences estimation is usually assessed by a test of the null hypothesis that the difference between the average outcomes of both groups is constant over time before the treatment. However, failure to reject the null hypothesis does not imply the absence of differences in time trends between both groups. We provide equivalence tests that allow researchers to find evidence in favor of the parallel trends assumption and thus increase the credibility of their treatment effect estimates. Since our test procedures are based on simple linear regressions, we show that they can be easily adapted to staggered treatment assignments and heterogeneous treatment effects by appropriate extensions of the regression model.

1 Introduction

In the classic case, the Difference-in-Differences (DiD) framework consists of two groups observed over two periods of time, where the "treatment group" is untreated in the initial period and has received a treatment in the second period whereas the "control group" is untreated in both periods. The key condition under which the DiD estimator yields sensible point estimates of the true effect of the treatment is known as the "parallel paths" or "parallel trends assumption", henceforth referred to as PTA, which states that in the absence of treatment both groups would have experienced the same temporal trends in the outcome variable on average. If pre-treatment observations are available for both groups, the plausibility of this assumption is typically assessed by plots accompanied by a formal testing procedure showing that there is no evidence of differences in trends over time between the treatment and the control group. However, this procedure is problematic as traditional pre-tests suffer from low power to detect violations of the PTA (Kahn-Lang & Lang 2020). Thus, finding no evidence of differences in trends in finite samples does not imply that there are no differences in trends in the population. More concerningly, Roth (2022) points out that if differences in trends exist, conditional on not detecting violations of parallel trends at the pre-testing stage, the bias of DiD-estimators may be greatly amplified.

Given the severe consequences of falsely accepting the PTA, we propose that instead of testing the null hypothesis of "no differences in trends" between the treatment and the control group in the pre-treatments periods, one should apply a test for statistical "equivalence". We provide three distinct types of equivalence that impose bounds on the maximum, the average and the root mean square change over time in the group mean difference between treatment and control in the pre-treatment periods. Given a threshold below which deviations from the PTA can be considered negligible, these tests allow the researcher to provide statistical evidence in favor of the PTA, thus increasing its credibility. If no sensible equivalence threshold can be determined before analyzing the data, we propose to report the smallest equivalence threshold for which the null hypothesis of "non-negligible trend differences" can still be rejected at a given level of significance. Conceptually, this idea is similar to the "equivalence confidence interval" in Hartman & Hidalgo (2018) applied to a DiD setting. Our procedure reverses the burden of proof since the data has to provide evidence *in favor* of similar trends in the treatment and the control group, which is arguably more appropriate for an assumption as crucial to the DiD-framework as the comparability of treatment and control in the absence of treatment. Furthermore, for our procedure based on equivalence tests, the power to reject the null hypothesis of a difference is increasing with the sample size (also see Hartman & Hidalgo 2018). This improves upon the current practice of testing the null hypothesis of "no difference", since large samples increase the chances of rejecting this null hypothesis (and thus seemingly making the DiD framework inapplicable), even if the true difference between treatment and control may be negligible in the given context. Finally, our equivalence test statistics make use of the standard OLS estimator and can thus easily be implemented in practice. This also allows us to extend our framework to heterogeneous treatment effects. For instance, we demonstrate how our tests can be applied in situations where treatment timing differs across groups (e.g. "staggered treatment assignment") or where average treatment effects depend on some observable characteristics.

As we use equivalence tests, our paper is closely related to Bilinski & Hatfield (2020), who provide a discussion on the benefits of using equivalence (or ”non-inferiority") tests when testing for violations of modeling assumptions. Their "one-step-up" approach is based on a non-inferiority test of treatment effect estimates obtained from a standard DiD model and from a model augmented with a particular violation of the parallel trends assumption (e.g. a linear trend). While both approaches stress the potential benefits of equivalence testing in DiD setups, a distinctive feature is that we do not necessarily focus on a particular violation of the PTA. As pointed out in Kahn-Lang & Lang (2020), including for instance group-specific linear time trends can lead to a loss in degrees of freedom and thus to a substantial loss in power. In contrast, our approach focuses on testing for "negligible" differences between treatment and control in the pre-treatment periods. Our paper is also related to other approaches that allow for certain deviations from exactly parallel trends. In particular, Rambachan & Roth (2022) relax the PTA by imposing restrictions on the extend in which post-treatment violations of parallel trends differ from pre-treatment differences in trends. They then proceed by deriving confidence sets that allow for uniformly valid inference when the imposed restrictions on trend differences are satisfied. Consequently, the parameters of interest are typically set-identified in their setup, where the identified set reflects the uncertainty about the PTA in the pre-treatment periods. In contrast, our approach focuses less on valid inference under violations of parallel trends but rather on testing for "negligible" differences in trends that are consistent with the pre-treatment data. A conceptual key difference is thus how the plausibility of point-identification based on the PTA is addressed: as the PTA imposes restrictions on counterfactuals and is thus inherently untestable, we take the point of view that one can point-identify the parameter of interest, as long as sufficient evidence in favor of the PTA is available from pre-treatment data, whereas Rambachan & Roth (2022) incorporate uncertainty about the PTA in their confidence intervals.

The rest of the paper is organized as follows. Section 2 introduces the main model and discusses the widely used practice of testing for violations of the PTA. Our equivalence tests as well as our main assumptions and theorems are presented in Section 3. Section 4 discusses the use of our methodology under violations of the PTA. Section 5 presents extensions of the main model that allow for heterogeneous treatment effects due to differences in treatment timing or observable characteristics. Simulation evidence on the performance of our test procedures is provided in Section 6, while Section 7 contains an empirical illustration of our approach. Section 8 concludes. Finally, mathematical details and tables are collected in the Appendix.

2 Pre-testing in Difference-in-Differences

To motivate our test procedures, we initially consider the simple DiD case with only two groups, homogeneous treatment effects and common treatment timing in a repeated cross-sectional setting where we observe $n_t \in \mathbb{N}$ individuals in period $t \in \{1, \ldots, T+1\}$. In later sections we extend our approach to allow for heterogeneous treatment effects and panel data models. We refer to individual i as "treated" or being in the "treatment group" if the treatment indicator $G_i = 1$ and as being "non-treated" or in the "control group" if $G_i = 0$. Moreover, periods $1, ..., T$ correspond to pre-treatment periods while $T+1$ denotes the post-treatment period. The potential outcome of unit i when treated is denoted as

Y_i^1, whereas Y_i^0 denotes the potential outcome of unit i in the absence of treatment. The observed outcome is then given by

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0)G_i \times D_{i,T+1}, \tag{2.1}$$

where $D_{i,l}$ denotes an indicator that takes the value 1 if unit i is observed in period $l \in \{1, ..., T+1\}$ and zero otherwise. Note that (2.1) implicitly imposes a "no-anticipation" assumption, as the observed outcome in the pre-treatment periods coincides with the potential outcome in the absence of treatment, which rules out any treatment effects before period $T+1$. Similar assumptions are used for instance in Goodman-Bacon (2021). Our object of interest is the average treatment effect on the treated

$$\pi_{ATT} := \mathbb{E}[Y_i^1 - Y_i^0 | G_i = 1, D_{i,T+1} = 1]. \tag{2.2}$$

The PTA, which ensures that in the absence of treatment both the treatment and the control group would have experienced the same time trends between the post-treatment period $T+1$ and period T (subsequently called the "base period" following Kahn-Lang & Lang 2020), is given by $\Delta_{T+1}(0) - \Delta_T(0) = 0$, where

$$\Delta_l(0) := \mathbb{E}[Y_i^0 | G_i = 1, D_{i,l} = 1] - \mathbb{E}[Y_i^0 | G_i = 0, D_{i,l} = 1],\ l = 1, ..., T+1\,.$$

In most applications, it is however not considered plausible that group trends are parallel between periods T and $T+1$ but not between period $l \in \{1, ..., T-1\}$ and T. In the rest of the paper, we therefore refer to the PTA in its "augmented" version given by

$$\Delta_l(0) - \Delta_T(0) = 0,\ l = 1, ..., T-1. \tag{2.3}$$

Under (2.1) and (2.3), we can recover the ATT as $\pi_{ATT} = \Delta_{T+1} - \Delta_T$, where

$$\Delta_l := \mathbb{E}[Y_i | G_i = 1, D_{i,l} = 1] - \mathbb{E}[Y_i | G_i = 0, D_{i,l} = 1]$$

denotes the population group mean difference in period l. Thus, assuming (2.1) holds, the PTA ensures that π_{ATT} can be estimated based on observable quantities.

2.1 Statistical evidence

A popular model specification (see, e.g., Angrist & Pischke 2008, p.177) that yields both an estimator of the ATT and a pre-testing procedure is

$$Y_i = c + \alpha G_i + \sum_{\substack{l=1 \\ l \neq T}}^{T+1} \gamma_l D_{i,l} + \sum_{\substack{l=1 \\ l \neq T}}^{T+1} \beta_l D_{i,l} \times G_i + u_i \ , \ i = 1, \ldots, n \ , \tag{2.4}$$

where c denotes a constant. A simple linear regression then yields estimates $\hat{\beta}_l$, $l \in \{1, \ldots, T-1, T+1\}$, where π_{ATT} is estimated by $\hat{\beta}_{T+1}$. The remaining $\hat{\beta}_0, \ldots, \hat{\beta}_{T-1}$ referring to leads of the treatment effect are used for a "Granger-type causality test" (Wing et al. 2018). If the trends in the average outcome of interest in treatment and control are indeed "parallel", changes in treatment status occurring in period $T+1$ should not affect the outcome in prior periods. Under strict exogeneity, i.e. $\mathbb{E}[u_i | G_i, D_{i,1}, ..., D_{i,T-1}, D_{i,T+1}] = 0$, we have $\beta_l = \Delta_l - \Delta_T$, i.e. β_l measures the change in group mean differences between period l and the base period. Thus, $\beta_l = 0$ signifies the absence of temporary shocks in periods l and T that only affect either treatment or control. Further notice that by (2.1) and (2.3), $\beta_l = \Delta_l(0) - \Delta_T(0)$=0, which underlines that anticipation of treatment is ruled out. Conversely, $\beta_l \neq 0$ signals that the control group may not be an optimal comparison group for the treatment group, as there may be unobserved differences between both. In that sense, $\beta_1, ..., \beta_{T-1}$ may provide a measure of comparability of treatment and control. To find evidence *against* the plausibility of parallel trends, it is therefore common in applied economic research to test for individual significance (see Roth 2022), i.e. for every $l \in \{1, ..., T-1\}$ we test

$$\mathrm{H}_0 : \ \beta_l = 0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_l \neq 0. \tag{2.5}$$

If the null hypothesis is rejected in a pre-treatment period, the PTA is deemed unreasonable, and consequently the DiD framework is often regarded as unsuitable in the corresponding context. This procedure has several shortcomings. For instance, a problematic common practice is to treat failure to reject the null hypothesis in (2.5) as evidence *in favor* of H_0, i.e. one proceeds as if the null hypothesis was true and as if the PTA held. From a statistical point of view, this practice is incorrect as it neglects the error of type II. In some cases, there may be differences in trends between both groups in the population that cannot be detected with traditional test of (2.5) due to a lack of statistical power. Roth (2022) points out that ignoring these differences can amplify the bias and thus raise concerns of a "publication bias", since articles using a DiD identification argument are more likely to be deemed publishable when a test of (2.5) could not detect evidence against the PTA. Moreover, the DiD framework is sometimes used even when H_0 in (2.5) is rejected, as some statistically significant differences are deemed negligible in a given context. However, a potential threshold $\mathcal{U} > 0$ that quantifies what constitutes a negligible effect is usually insufficiently discussed. On the other hand, if the DiD framework is not applied when H_0 in (2.5) is rejected in at least one pre-treatment period, useful information may be lost if $\mathcal{U}$ can be interpreted as a plausible "upper bound" for trend differences. For these reasons, the plausibility of the PTA as the fundamental modeling assumption of the DiD framework can be more convincingly assessed using statistical equivalence tests as these tests address all of the above shortcomings of the current standard testing procedure. For instance, to rewrite (2.5) in terms of statistical equivalence, for some $l \in \{1, ..., T-1\}$ one would define the equivalence threshold $\mathcal{U} > 0$ and test

$$\mathrm{H}_0: \ |\beta_l| \geq \mathcal{U} \qquad \text{vs.} \qquad \mathrm{H}_1: |\beta_l| < \mathcal{U}.$$

Rejecting this null hypothesis thus yields evidence *in favor* of the absence of changes in group mean differences between period t and the base period. In the following, we elaborate on the benefits of equivalence tests and provide different ways of summarizing the statistical evidence in favor of the PTA in the pre-treatment periods.

3 Testing for equivalence

Equivalence testing is well known in biostatistics (see Berger & Hsu 1996 or Wellek 2010). While it has recently been considered in the statistical literature for the analysis of structural breaks (e.g. Dette & Wied 2014, Dette & Wu 2019, Dette, Kokot & Aue 2020 or Dette, Kokot & Volgushev 2020), it is less frequently used in econometrics. Instead of assuming that treatment and control are perfectly comparable (i.e. $\beta_l = 0$, $l = 1, ..., T-1$) unless there is strong evidence *against* this assumption, we suggest several testing procedures that explicitly require finding evidence *in favor* of the comparability of both groups. Each of the tests is based on an upper bound $\mathcal{U} > 0$ for changes in the group mean differences in the pre-treatment periods relative to the base period. There are two ways in which one can make use of the upper bound $\mathcal{U}$. First, as in the "classic" use of equivalence tests, one can specify a threshold $\mathcal{U}$ below which changes in the group mean differences over time are deemed negligible. One then applies our equivalence testing procedures using the pre-specified threshold. Rejecting an equivalence test with threshold $\mathcal{U}$ at α level of significance then implies that deviations from parallel trends in the pre-treatment periods are negligible with probability $1 - \alpha$. Since the PTA in the pre-treatment periods is now supported by sufficient evidence, this provides a justification for the PTA post-treatment so that the true ATT can again be point-identified. This procedure improves upon the current use of the Granger-causality test as it requires an explicit rationalization of the threshold $\mathcal{U}$ and sufficient data to support the assumption of negligible violations of the PTA pre-treatment. The choice of the threshold $\mathcal{U}$ should reflect the specific scientific background of the application. In bio-statistics, the popularity of equivalence tests has led to a consensus on sensible choices for $\mathcal{U}$, and regulators frequently specify the equivalence thresholds that should be employed (see Wellek 2010 for a recent review). We expect that with a more frequent adoption of equivalence testing in applied economics a similar consensus will be reached. However, in some applications, it may still be difficult to objectively argue that a certain extend of violations of the PTA can be ignored in practice. It may then be sensible to report $\mathcal{U}^*$ as the smallest value at which H_0 can be rejected at a given level of significance (i.e. for which "equivalence of pre-trends" can be concluded). Small values of $\mathcal{U}^*$ relative to the estimated treatment effect may then be regarded as reassuring as it is unlikely that the treatment effect is merely an artifact of differences in trends. On the contrary, if $\mathcal{U}^*$ is relatively large, the credibility of the estimated effect is in serious doubt. Finally, in cases where the choice of the threshold is difficult, the methodology presented here can also be used to provide (asymptotic) confidence intervals for violations of the PTA which provide information about the size of the deviation with statistical guarantees. Notice that these confidence intervals differ from those considered by Rambachan & Roth (2022), as they focus on confidence intervals for the treatment effect that reflect the uncertainty about the PTA while we consider confidence intervals on trend differences in the pre-treatment periods. Consequently, unlike Rambachan & Roth (2022), we maintain point-identification of the ATT in the presence of sufficient evidence in favor of the PTA.

Overall, we consider three distinct hypotheses to test for equivalence of pre-trends in treatment and control. We start with a discussion of the maximum absolute change of the group mean difference in the pre-treatment periods relative to the base period. More precisely, letting $\beta^{(T-1)} := (\beta_1, ..., \beta_{T-1})'$, for a given level of significance α and the equivalence threshold $\delta > 0$, we test

$$\mathrm{H}_0 : \|\beta^{(T-1)}\|_\infty \geq \delta \qquad \text{vs.} \qquad \mathrm{H}_1 : \|\beta^{(T-1)}\|_\infty < \delta, \tag{3.1}$$

where $\|\beta^{(T-1)}\|_\infty := \max_{l\in\{1,...,T-1\}} |\beta_l|$. Since we are now controlling the type I error, this implies that with probability of at least $1-\alpha$, δ is an upper bound for the absolute change in group mean differences in the pre-treatment periods relative to the base period.

In many applications, pre- and post-treatment periods are pooled, for instance to increase statistical power. Similarly, it may be sensible in some applications to consider a pooled or average measure of the pre-treatment deviations from parallel trends. Thus, defining $\bar{\beta}^{(T-1)} := \frac{1}{T-1}\sum_{l=1}^{T-1}\beta_l$, one can find bounds on the average deviation from the group mean difference in the base period by testing

$$\mathrm{H}_0 : |\bar{\beta}^{(T-1)}| \geq \tau \qquad \text{vs.} \qquad \mathrm{H}_1 : |\bar{\beta}^{(T-1)}| < \tau. \tag{3.2}$$

One disadvantage of (3.2) is that there may be cancellation effects in situations where the components of $\beta^{(T-1)}$ are large in absolute terms but have opposing signs. Therefore, (3.2) should be used when differences in pre-trends can safely assumed to be of the same sign. As pointed out in Rambachan & Roth (2022), monotone violations of the PTA are frequently discussed in the applied literature. For instance, treatment effect estimates are often considered robust if potential violations of the PTA are of the opposing sign and can thus be ruled out as an explanation for the estimated effects. As an alternative to (3.2) that does not suffer from potential cancellation effects, we further consider the root mean square (RMS) of $\beta^{(T-1)}$, i.e. $\beta_{RMS} := \|\beta^{(T-1)}\|/\sqrt{T-1} = \sqrt{\frac{1}{T-1}\sum_{l=1}^{T-1}\beta_l^2}$, where $\|\cdot\|$ denotes the euclidean norm on $\mathbb{R}^{T-1}$. The RMS of $\beta^{(T-1)}$ can thus be interpreted as the euclidean distance between treatment and control in the pre-treatment periods relative to the distance in the base period scaled by the number of pre-treatment periods. The scaling is induced to ensure that this distance between treatment and control does not increase with the number of pre-treatment periods available. The hypotheses are then formulated as

$$\mathrm{H}_0 : \beta_{RMS} \geq \zeta \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_{RMS} < \zeta \ , \tag{3.3}$$

which can equivalently be written as

$$\mathrm{H}_0 : \beta_{RMS}^2 \geq \zeta^2 \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_{RMS}^2 < \zeta^2. \tag{3.4}$$

In Section 3.1 below we develop a test statistic for (3.4) and recover ζ as $\sqrt{\zeta^2}$.

3.1 Implementing equivalence tests

We now focus on developing the test statistics for the hypotheses in (3.1), (3.2) and (3.4) which can be applied in model (2.4). To formalize the necessary assumptions, we introduce the random vector

$$W_i := \big(1, G_i, D_{i,1}, \dots, D_{i,T-1}, D_{i,T+1}, G_i \times D_{i,1}, \dots, G_i \times D_{i,T-1}, G_i \times D_{i,T+1}\big)'$$

and the parameter

$$\theta := (c, \alpha, \gamma_1, ..., \gamma_{T-1}, \gamma_{T+1}, \beta_1, ..., \beta_{T-1}, \beta_{T+1})' \in \mathbb{R}^{2T+2}. \tag{3.5}$$

With these notations we can write model (2.4) in the form $Y_i = W_i'\theta + u_i$, and the least squares estimator $\hat{\theta}$ is given by

$$\hat{\theta} = (\frac{1}{n}\sum_{i=1}^{n} W_i W_i')^{-1} \frac{1}{n}\sum_{i=1}^{n} W_i Y_i = \theta + (\frac{1}{n}\sum_{i=1}^{n} W_i W_i')^{-1} \frac{1}{n}\sum_{i=1}^{n} W_i u_i, \tag{3.6}$$

where $n := \sum_{t=1}^{T+1} n_t$ denotes the total sample size. For the asymptotic analysis we make the following assumptions.

Assumption 3.1

(1) G_i is a Bernoulli distributed random variable with parameter $p \in (0, 1)$ specifying the probability of individual i being treated.

(2) The $T+1$-dimensional vector $(D_{i,1}, \dots, D_{i,T+1})'$ has a multinomial distribution with a single trial and probabilities $p_1, \dots, p_{T+1}$, where $p_j \in (0,1)$ specifies the probability that individual i is observed in period j and $\sum_{j=1}^{T+1} p_j = 1$.

(3) $W_1, \dots, W_n$ and $u_1 \dots, u_n$ are independent samples of independent identically distributed random variables.

(4) The matrix $\Gamma = \mathbb{E}[W_i W_i']$ exists and is positive definite. $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[u_i^2]$ exists and is positive.

Under these assumptions, standard arguments show that the estimate $\hat{\theta}$ is consistent for θ in (3.5). Let further $\hat{\beta} := (\hat{\beta}_1, ..., \hat{\beta}_{T-1}, \hat{\beta}_{T+1})'$ denote the OLS estimator of the parameter $\beta := (\beta_1, ..., \beta_{T-1}, \beta_{T+1})'$ in model (2.4). It then follows that

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{N}(0, \Sigma), \tag{3.7}$$

where $\mathrm{N}(0, \Sigma)$ denotes a T-dimensional normal distribution with mean vector $0 \in \mathbb{R}^T$ and covariance matrix $\Sigma = (\Sigma_{ij})_{i,j=1,\dots,T}$. As we discuss in Remark 3.2 below, our methodology also works under alternative assumptions which for instance allow for serial dependence in the model errors or panel data applications. Combining (2.1) and (2.4) with Assumption 3.1 now implies that $\beta = (0, \dots, 0, \pi_{ATT})$ if and only if the PTA is satisfied. Based on the asymptotic normality of the OLS estimator in (3.7), we propose tests for the three different hypotheses of equivalence.

3.1.1 Two tests for (3.1).

To describe the first test for the hypotheses in (3.1) we initially consider the case $T = 2$ so that our objective is to test whether a single parameter β_1 exceeds a certain threshold. As $\hat{\beta}_1$ is approximately distributed as $\mathrm{N}_1(\beta_1, \Sigma_{11}/n)$, the test statistic $|\hat{\beta}_1|$ approximately follows a folded normal distribution. We therefore propose to reject the null hypothesis in (3.1), whenever

$$|\hat{\beta}_1| < \mathcal{Q}_{\mathrm{N}_F(\delta, \hat{\Sigma}_{11}/n)}(\alpha), \tag{3.8}$$

where $\mathcal{Q}_{\mathrm{N}_F(\delta,\sigma^2)}(\alpha)$ denotes the α quantile of the folded normal distribution with mean δ and variance σ^2 and where $\hat{\Sigma} = (\hat{\Sigma}_{ij})_{i,j=1,\dots,T}$ is a consistent estimator of the matrix Σ in

(3.7). It is shown in Appendix A that this test is consistent, has asymptotic level α and is (asymptotically) uniformly most powerful for testing the hypotheses in (3.1) in the case $T = 2$. In particular this test is more powerful than the two-sided t-test (TOST), which could be developed following the arguments in Hartman & Hidalgo (2018). For $T > 2$, we apply the idea of intersection-union tests outlined in Berger & Hsu (1996) and reject the null hypothesis in (3.1), whenever

$$|\hat{\beta}_t| < \mathcal{Q}_{\mathrm{N}_F(\delta,\hat{\Sigma}_{tt}/n)}(\alpha)\ \forall t \in \{1, \dots, T-1\}. \tag{3.9}$$

While this test is computationally attractive, a well-known disadvantage of testing procedures based on the intersection-union principle is that they tend to be rather conservative (see Berger & Hsu 1996, among others), which is confirmed by our simulation study (see Table 2 in Section 6).

To obtain a more powerful test for the hypotheses (3.1), we write $\hat{\beta}^{(T-1)} := (\hat{\beta}_1, \dots, \hat{\beta}_{T-1})'$ so that $\hat{\beta}^{(T-1)}$ denotes the sub-vector which extracts the coordinates in the positions $T+3, \dots, 2T$ from the vector $\hat{\theta}$. We derive an alternative test for (3.1) as follows: In the first step, estimate (2.4) by OLS to obtain the unconstrained least squares estimator $\hat{\theta}_u$ and the sub-vector $\hat{\beta}_u^{(T-1)}$. In the second step, we re-estimate (2.4) by minimizing the sum of squared residuals under the constraint $\max_{l=1,\dots,T-1} |\beta_l| = \delta$ to obtain a constrained estimator, say $\hat{\theta}_c$. We then define new estimators of the parameters as

$$\hat{\hat{\theta}}_c = \begin{cases} \hat{\theta}_u & \text{if} \quad \|\hat{\beta}^{(T-1)}\|_\infty \geq \delta \\ \hat{\theta}_c & \text{if} \quad \|\hat{\beta}^{(T-1)}\|_\infty < \delta \end{cases} \tag{3.10}$$

and $\hat{\hat{\sigma}}_c = \frac{1}{n-2T-2} \sum_{i=1}^n (Y_i - W_i' \hat{\hat{\theta}}_c)^2$. Note that the vector $\hat{\hat{\beta}}_c$ extracted from $\hat{\hat{\theta}}_c$ satisfies the null hypothesis in (3.1). In the third step, for $b = 1, ..., B \in \mathbb{N}$, we generate bootstrap samples with $u_1^{(b)}, ..., u_n^{(b)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \hat{\hat{\sigma}}_c)$ and $Y_1^{(b)} = W_i' \hat{\hat{\theta}}_c + u_i^{(b)}$. For each bootstrap sample, estimate $\hat{\theta}_u^{(b)}$ and extract the components $\hat{\beta}_1^{(b)}, ..., \hat{\beta}_{T-1}^{(b)}$. Further compute $\mathcal{Q}_\alpha^*$ as the empirical α-quantile of the bootstrap sample $\{\max_{l=1,\dots,T-1} |\hat{\beta}_l^{(b)}| : b = 1, ..., B\}$. Finally, reject the null hypothesis H_0 in (3.1) if

$$\|\hat{\beta}^{(T-1)}\|_\infty < \mathcal{Q}_\alpha^* \ . \tag{3.11}$$

The following result shows that this test is consistent and has asymptotic level α.

Theorem 3.1 *The test defined by* (3.11) *is consistent and has asymptotic level* α *for the hypotheses in* (3.1)*. More precisely,*

(1) if the null hypothesis in (3.1) *is satisfied, then we have for any* $\alpha \in (0, 0.5)$

$$\limsup_{n\to\infty} \mathbb{P}_{\beta^{(T-1)}}\big(\|\hat{\beta}^{(T-1)}\|_\infty < \mathcal{Q}_\alpha^*\big) \leq \alpha. \tag{3.12}$$

(2) if the null hypothesis in (3.1) *is satisfied and the set*

$$\mathcal{E} = \{\ell = 1, \ldots, T-1 \ : \ |\beta_\ell| = \|\beta^{(T-1)}\|_\infty\} \tag{3.13}$$

consists of one point, then we have for any $\alpha \in (0, 0.5)$

$$\lim_{n\to\infty} \mathbb{P}_{\beta^{(T-1)}}\big(\|\hat{\beta}^{(T-1)}\|_\infty < \mathcal{Q}_\alpha^*\big) = \begin{cases} 0 & \textit{if} \ \ \|\beta^{(T-1)}\|_\infty > \delta \\ \alpha & \textit{if} \ \ \|\beta^{(T-1)}\|_\infty = \delta. \end{cases} \tag{3.14}$$

(3) if the alternative in (3.1) *is satisfied, then we have for any* $\alpha \in (0, 0.5)$

$$\lim_{n\to\infty} \mathbb{P}_{\beta^{(T-1)}}\big(\|\hat{\beta}^{(T-1)}\|_\infty < \mathcal{Q}_\alpha^*\big) = 1. \tag{3.15}$$

3.1.2 A test for (3.2).

For some fixed $\tau > 0$, a test can be constructed by first computing the statistic

$$\bar{\hat{\beta}}^{(T-1)} := \frac{1}{T-1}\sum_{t=1}^{T-1} \hat{\beta}_t = \mathbb{1}'\hat{\beta}^{(T-1)}/(T-1),$$

where $\mathbb{1} = (1, \ldots, 1)' \in \mathbb{R}^{T-1}$. Note that it follows from (3.7) that

$$\sqrt{n}\mathbb{1}'(\hat{\beta}^{(T-1)} - \beta^{(T-1)}) \to \mathrm{N}(0, \mathbb{1}'\Sigma\mathbb{1})).$$

Consequently, based on the discussion in the first part of 3.1.1, we propose to reject the null hypothesis in (3.2), whenever

$$|\bar{\hat{\beta}}^{(T-1)}| < \mathcal{Q}_{\mathrm{N}_F(\tau, \hat{\sigma}^2)}(\alpha), \tag{3.16}$$

where $\hat{\sigma}^2 = \mathbb{1}'\hat{\Sigma}\mathbb{1}/(n(T-1)^2)$.

3.1.3 A pivotal test for (3.4).

In order to construct a pivot test for the hypotheses (3.4), recall the definition of the OLS estimator $\hat{\theta}$ in (3.6) and let $\varepsilon > 0$ denote a small positive constant. For $\lambda \in [\varepsilon, 1]$, define

$$\hat{\theta}(\lambda) = \Big(\frac{1}{n}\sum_{i=1}^{\lfloor n\lambda \rfloor} W_i W_i'\Big)^{-1} \frac{1}{n}\sum_{i=1}^{\lfloor n\lambda \rfloor} W_i Y_i$$

as the OLS estimator for θ in (3.5) from the sample $(W_1, Y_1), \dots, (W_{\lfloor n\lambda \rfloor}, Y_{\lfloor n\lambda \rfloor})$, such that for sufficiently large sample sizes $\hat{\theta}(\lambda)$ is well defined. Next, define $\hat{\beta}^{(T-1)}(\lambda)$ as the subvector of $\hat{\theta}(\lambda)$ extracting the coordinates in the positions $T+2, \dots, 2T+2$. Further define $\hat{\beta}^2_{RMS} := \frac{1}{T-1}\|\hat{\beta}^{(T-1)}\|^2$ and $\hat{\beta}^2_{RMS}(\lambda) := \frac{1}{T-1}\|\hat{\beta}^{(T-1)}(\lambda)\|^2$. Notice that for $\lambda = 1$ we recover the respective estimators based on the full sample. We now define

$$\hat{M}_n := \frac{\hat{\beta}^2_{RMS}(1) - \beta^2_{RMS}}{\hat{V}_n}, \tag{3.17}$$

where

$$\hat{V}_n = \Big(\int_{\varepsilon}^{1} (\hat{\beta}^2_{RMS}(\lambda) - \hat{\beta}^2_{RMS}(1))^2 \nu(d\,\lambda)\Big)^{1/2} \tag{3.18}$$

and ν denotes a measure on the interval $[\varepsilon, 1]$. The following result is proved in the Appendix.

Theorem 3.2 *If Assumption 3.1 is satisfied and $\beta^{(T-1)} \neq 0$, then the statistic $\hat{M}_n$ defined in* (3.17) *converges weakly with a non-degenerate limit distribution, that is*

$$\hat{M}_n \xrightarrow{d} \mathbb{W} := \frac{\mathbb{B}(1)}{\big(\int_{\varepsilon}^{1}(\mathbb{B}(\lambda)/\lambda - \mathbb{B}(1))^2\nu(d\lambda)\big)^{1/2}}, \tag{3.19}$$

where $\{\mathbb{B}(\lambda)\}_{\lambda\in[\varepsilon,1]}$ is a Brownian motion on the interval $[\varepsilon, 1]$. Moreover, if $\beta^{(T-1)} = 0$, then

$$\hat{M}_n \xrightarrow{d} \frac{\mathbb{Z}^2(1)}{\big(\int_{\varepsilon}^{1}(\mathbb{Z}^2(\lambda) - \mathbb{Z}^2(1))^2\nu(d\lambda)\big)^{1/2}}\ , \tag{3.20}$$

where $\mathbb{Z}^2(\lambda) = \frac{1}{\lambda^2}\vec{\mathbb{B}}'(\lambda)D'D\vec{\mathbb{B}}$, $\approx\eqsim\vec{\gtrdot}\ltimes\approxeq$ is a $2T+2$-dimensional vector of independent Brownian motions and D is a $(T-1)\times(2T+2)$ matrix of full rank defined in equation (A.7) *in the Appendix.*

It follows from the proof of Theorem 3.2 that the statistic $\hat{\beta}^2_{RMS}$ is a consistent estimator of β^2_{RMS}. Therefore, we propose to reject the null hypothesis H_0 in (3.4) (and consequently H_0 in (3.3)), whenever

$$\hat{\beta}^2_{RMS} < \zeta^2 + \mathcal{Q}_{\mathbb{W}}(\alpha)\hat{V}_n \tag{3.21}$$

where $\mathcal{Q}_{\mathbb{W}}(\alpha)$ is the α-quantile of the distribution of the limiting distribution of the random variable $\mathbb{W}$ on the right-hand side of (3.19). Note that these quantiles can be easily obtained by simulation because the distribution of $\mathbb{W}$ is completely known. For instance, $\mathcal{Q}_{\mathbb{W}}(0.05) \approx -2.1$. The following result shows that this decision rule defines a valid test for the hypotheses in (3.4).

Theorem 3.3 *If Assumption 3.1 is satisfied, then the test defined by* (3.21) *is a consistent asymptotic level α-test for the hypotheses in* (3.4)*, that is*

$$\lim_{n\to\infty} \mathbb{P}_{\beta^{(T-1)}}\Big(\hat{\beta}^2_{RMS} < \zeta^2 + \mathcal{Q}_{\mathbb{W}}(\alpha)\hat{V}_n\Big) = \begin{cases} 0, & \text{if} \quad \beta^2_{RMS} > \zeta^2 \\ \alpha, & \text{if} \quad \beta^2_{RMS} = \zeta^2 \\ 1, & \text{if} \quad \beta^2_{RMS} < \zeta^2 \end{cases}$$

Remark 3.1

(a) Notice that in practice one chooses ν as a discrete distribution which makes the evaluation of the integrals in (3.18) and in the denominator of the random variable $\mathbb{W}$ very easy. For example, if ν denotes the uniform distribution on $\{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\}$, then the statistics $\hat{V}^2_n$ in (3.18) simplifies to

$$\frac{1}{4}\sum_{k=1}^{4}\Big(\big\|\hat{\beta}^{(T-1)}(\tfrac{k}{5})\big\|^2 - \big\|\hat{\beta}^{(T-1)}(1)\big\|^2\Big)^2 .$$

This measure is also used in the simulation study in Section 6, where we analyze the finite sample properties of the different procedures. In practice, it is thus not necessary to explicitly choose ε.

(b) It follows from the proof of Theorem 3.2 that an asymptotic $(1-\alpha)$-confidence interval for the parameter $\beta^2_{RMS} > 0$ is given by

$$\Big[\hat{\beta}^2_{RMS} + \mathcal{Q}_{\mathbb{W}}(\alpha/2)\hat{V}_n,\ \hat{\beta}^2_{RMS} + \mathcal{Q}_{\mathbb{W}}(1-\alpha/2)\hat{V}_n\Big].$$

(c) Theorem 3.3 can be extended to get uniform results. More precisely, define for a small positive constant c the sets

$$\mathcal{H} = \Big\{ \beta^{(T-1)} \; \Big| \; \Delta(\beta^{(T-1)}) > c \ , \ \hat{\beta}^2_{RMS} \geq \zeta^2 \Big\} \ ,$$
$$\mathcal{A} = \Big\{ \beta^{(T-1)} \; \Big| \; \Delta(\beta^{(T-1)}) > c \ , \ \hat{\beta}^2_{RMS} < \zeta^2 - x/\sqrt{n} \Big\} \ ,$$

corresponding to the null hypothesis and alternative, respectively, where $\Delta(\beta^{(T-1)})$ is defined in equation (A.8) in the Appendix. Then

$$\limsup_{n\to\infty} \sup_{\beta^{(T-1)}\in\mathcal{H}} \mathbb{P}_{\beta^{(T-1)}}\Big(\hat{\beta}^2_{RMS} < \zeta^2 + \mathcal{Q}_{\mathbb{W}}(\alpha)\hat{V}_n \Big) = \alpha \ .$$

Furthermore there exists a non-decreasing function $f : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, with $f(x) > \alpha$ for all $x > 0$ and $\lim_{x\to\infty} f(x) = 1$, such that

$$\liminf_{n\to\infty} \inf_{\beta^{(T-1)}\in\mathcal{A}} \mathbb{P}_{\beta^{(T-1)}}\Big(\hat{\beta}^2_{RMS} < \zeta^2 + \mathcal{Q}_{\mathbb{W}}(\alpha)\hat{V}_n \Big) = f(x) \ .$$

The details are omitted for the sake of brevity.

Remark 3.2 The statements made in this section remain valid under more general or alternative assumptions and we exemplary mention here two such scenarios.

(1) In Assumption 3.1 it is postulated that the random variables $(W_1, \varepsilon_1), \ldots, (W_n, \varepsilon_n)$ are independent. However, a careful inspection of the proofs in Section A.3 shows that similar results can be obtained in the case of dependent data. More precisely, for a symmetric $d \times d$ matrix A let vech(A) denote the $d(d+1)/2$-dimensional vector that stacks the columns of the matrix A below the diagonal in a vector. Let $d := (2T+2) + (2T+2)(2T+3)/2$, let K denote a non-singular $d \times d$-matrix and let $\vec{\mathbb{B}}$ be a d-dimensional vector of independent Brownian motions. If the time series $\{(W_i, u_i)\}_{i=1}^n$ is stationary and the sequential process satisfies

$$\left\{ \sqrt{n} \begin{pmatrix} \frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} W_i u_i \\ \mathrm{vech}\big(\frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} W_i W_i' - \Gamma\big) \end{pmatrix} \right\}_{\lambda\in[\varepsilon,1]} \rightsquigarrow \Big\{ K \frac{\mathbb{B}(\lambda)}{\lambda} \Big\}_{\lambda\in[\varepsilon,1]}, \tag{3.22}$$

where "$\rightsquigarrow$" denotes weak convergence in the space $(\ell^\infty[\varepsilon,1])^d$ of all d-dimensional bounded functions on the interval $[\varepsilon,1]$, then the results stated in this section remain valid. Results of the form (3.22) have been proved for many dependence concepts in the literature (such as different types of mixing or physical dependence; see, for instance, Merlevède et al. 2006 and the references therein).

(2) Similarly, note that Assumption 3.1(2), which reflects the fact that each individual is only observed at exactly one time period, can be replaced by other assumptions, modeling alternative observation schemes. For example, in the situation of panel data with no missing observations, the vector $D_i = (D_{i,1}, \dots, D_{i,T+1})'$ is not random and given by $(1, \dots, 1)'$. Moreover, panel data with missing observations can be also modeled using a random vector $D_i = (U_{i,1}, \dots, U_{i,T+1})'$ where $U_{i,1}, \dots, U_{i,T+1}$ are independent Bernoulli variables with success probabilities $p_1, \dots, p_{T+1}$, respectively. Here, $1 - p_t$ represents the probability that an observation for the i-th individual is not available for time t.

4 Equivalence testing in practice

In practice, the PTA is often violated, for instance due to self-selection into treatment that is not accounted for in the estimation procedure (Heckman & Smith 1999). Importantly for our approach, violations of the PTA due to differences in unobserved characteristics between both groups typically induce bias in $\hat{\beta}^{(T-1)}$ which affects the equivalence thresholds at which the null hypothesis in our tests can be rejected. In order to formalize the violations of the PTA, we now consider the model in (2.4) with the crucial difference that the model error may contain a vector of unobserved covariates that lead to unobserved differences between the two groups, thus making the control group an imperfect comparison group for the treatment group. For instance, the variable Z_i may represent group-specific transitory shocks leading to a pre-program-dip or other unobserved individual characteristics (e.g. the sector of last employment) that affect the mean difference of the outcome of interest between the two groups. The data generating process is thus given by

$$Y_i = c + \alpha G_i + \sum_{\substack{l=1 \\ l \neq T}}^{T+1} \gamma_l D_{i,l} + \sum_{\substack{l=1 \\ l \neq T}}^{T+1} \beta_l \, D_{i,l} \times G_i + \tilde{u}_i, \qquad i = 1, \dots, n \tag{4.1}$$

where $\tilde{u}_i = Z_i' \nu + u_i$. When Z_i is omitted, the OLS estimator is

$$\hat{\theta} = \theta + (\frac{1}{n} \sum_{i=1}^{n} W_i W_i')^{-1} \frac{1}{n} \sum_{i=1}^{n} W_i Z_i' \nu + (\frac{1}{n} \sum_{i=1}^{n} W_i W_i')^{-1} \frac{1}{n} \sum_{i=1}^{n} W_i u_i.$$

As $\mathbb{E}[W_i u_i] = 0$ under Assumption 3.1, it is easy to see that the OLS estimator is only consistent as $n \to \infty$ if $\mathbb{E}[W_i Z_i'] = 0$. Notice that in the presence of Z_i we have $\beta_l =$

$\Delta_l - \Delta_T - (\Delta_l^Z - \Delta_T^Z)$, where

$$\Delta_l^Z := \mathbb{E}[Z_i|G_i = 1, D_{i,l} = 1] - \mathbb{E}[Z_i|G_i = 0, D_{i,l} = 1].$$

Thus, in the presence of unobserved covariates that affect the group means of treatment and control differently, the OLS estimator is biased and estimates $\theta + \rho$, where $\rho = \Gamma^{-1}\mathbb{E}[W_i Z_i' \nu]$ is the omitted variable bias. Therefore, when the true effect of the treatment prior to the treatment is zero, i.e. $\beta_l = 0$ for $l = 1, ..., T-1$, rejecting our equivalence tests for a threshold $\mathcal{U} > 0$ implicitly yields an upper bound for the omitted variable bias ρ. In the following, let δ^*, τ^* and ζ^* denote the smallest values such that the null hypotheses in (3.1), (3.2) and (3.3) can be rejected.

4.1 Examples

We now consider possible scenarios in which the PTA is violated due to the presence of unobserved covariates that have a differential effect on both groups. To simplify the exposition, we assume that the unobserved variable only affects the treatment group while the control group is unaffected.

Example 4.1 (Pre-program dip) As a first example, we model Ashenfelter's dip through the presence of a temporary shock denoted as Z_i that affects one group but not the other. We assume that the data is generated by the model in (4.1) with $Z_i = D_{i,T} \times G_i \times V_i$, where V_i, $i = 1, ..., n$, denotes i.i.d draws of a random variable with mean $v > 0$ and bounded variance independent of treatment status and time. We further assume that the treatment itself does not have an effect before the treatment takes place so that $\beta_1, ..., \beta_{T-1} = 0$. The OLS estimator of π_{ATT}, which still corresponds to the usual change in mean difference of the outcome variable from the post-treatment period to the base period then becomes

$$\Delta_{T+1} - \Delta_T = \beta_{T+1} + \Delta_{T+1}^Z - \Delta_T^Z = \beta_{T+1} - \nu,$$

since $\Delta_{T+1}^Z = 0$ and $\Delta_T^Z = \nu$. Therefore, we cannot recover the true ATT β_{T+1} due to the omitted variable bias $\rho = -\nu$. However, a similar argument shows that $\hat{\beta}_l$ converges to $\beta_l - v = -v$ for $l \in \{1, ..., T-1\}$ which differs from the true β_1 by the same amount in absolute terms as the probability limit of the estimated treatment effect differs from the true treatment effect. Consequently, if the null hypotheses in (3.1), (3.2) or (3.3) is rejected for a threshold $\mathcal{U}$ at level of significance α, $\mathcal{U}$ constitutes an upper bound of the absolute omitted variable bias with probability of at least $1-\alpha$.

Example 4.2 (Unobserved covariate with time trend) We now consider the DGP in (4.1) when the unobserved variable Z_i follows a time trend. More precisely, the unobserved variable is modeled as $Z_i = \psi \times G_i \times D_{i,l} \times l$, where ψ represents the slope of the time trend which only affects the treatment group and $l \in \{1, \ldots, T+1\}$. In this setup, $\Delta_{T+1} - \Delta_T = \beta_{T+1} + \psi$, since $\Delta^Z_{T+1} - \Delta^Z_T = (T+1)\psi - T\psi = \psi$. Moreover, for $l \in \{1, ..., T-1\}$, $\beta_l - \Delta_T = \beta_l + \psi(l-T) = \psi(l-T)$, so that $|\hat{\beta}_l|$ will typically increase with $|l-T|$. Thus, δ^*, τ^* and ζ^* will typically increase accordingly with the number of pre-treatment periods available. While δ^* increases with T even in the absence of an underlying time trend, the increase in τ^* and ζ^* can be regarded as evidence against the PTA and temporary shocks to the group mean difference (as in Ashenfelter's dip) and may thus be useful in identifying a permanent time trend.

5 Equivalence testing with heterogeneous treatment effects

The use of the simple DiD model has recently experienced substantial criticism in the presence of multiple groups, heterogeneous treatment effects and differences in treatment timing. In this situation, the DiD estimator often does not correspond to a reasonable estimate of the ATT (see, for instance, Goodman-Bacon 2021, Callaway & Sant'Anna 2021, Sun & Abraham 2021, Borusyak et al. 2021 or de Chaisemartin & D'Haultfœuille 2020. Excellent reviews of this fast-growing literature are provided by Roth et al. 2022 and de Chaisemartin & D'Haultfoeuille 2022.). In a recent paper, Wooldridge (2021) shows that this deficiency of the DiD estimator can be regarded as a model misspecification problem. He then proposes model adjustments that allow for treatment effect heterogeneity due to differences in treatment timing and observed characteristics (which are assumed to be unaffected by the treatment). In the following, we show how our equivalence tests can be adapted for the case of staggered adoption over time of an absorbing treatment in the presence of a never-treated group. Here, following Roth et al. (2022), we refer to a treatment as "staggered" if some groups are treated earlier than others. The treatment is "absorbing" if treated units remain treated in periods after the initial treatment assignment. Following (Wooldridge 2021, Section 6), we assume that the time since the initial treatment adoption produces different levels of exposure to the treatment, resulting in treatment effect heterogeneity across time. As before, we consider repeated cross-sections where each individual i is observed in exactly one period and treatment cohort (the panel data case can be handled by adjusting the notation as in Wooldridge 2021), and we maintain the assumption that T pre-treatment periods are observed. In each of the following periods $T+1,...,\overline{T}$, a subset of individuals adopts treatment, leading to "treatment cohorts". To define a treatment cohort dummy, let $G_i^r = 1$ if individual i has first adopted treatment in period $r \in \mathcal{R} := \{T+1,...,\overline{T},\infty\}$ and zero otherwise, where G_i^∞ is a dummy indicating that individual i is a member of the never treated group. We assume that individuals of every treatment cohort can be observed in each period, i.e. $\mathbb{P}(G_i^r \times D_{i,s} = 1) = p_{rs} \in (0,1)$ with $\sum_{r,s} p_{rs} = 1$. The potential outcome of unit i in treatment cohort $r \in \mathcal{R}$ observed in time period $t \in \{1,...,\overline{T}\}$ is denoted by $Y_i^r(t)$, where the "baseline" potential outcome in period t if unit i is never treated is given by $Y_i^\infty(t)$. We assume that the observed outcome can be written as

$$Y_i = \sum_{t=1}^{\overline{T}} Y_i^\infty(t) D_{i,t} + \sum_{r=T+1}^{\overline{T}} \sum_{t=r}^{\overline{T}} (Y_i^r(t) - Y_i^\infty(t)) D_{i,t} G_i^r \tag{5.1}$$

which implicitly rules out that units deviate from their designated treatment paths. In particular, (5.1) rules out anticipatory behavior or spillover effects. In this staggered setting, researchers may be interested in estimating the ATTs for each of the post-treatment periods separately. If these ATTs are dependent on a vector of observed covariates X_i, the ultimate objects of interest are

$$\pi_X^r(t) = \mathbb{E}[Y_i^r(t) - Y_i^\infty(t) | G_i^r = 1, D_{i,t} = 1, X_i], r = T+1,...,\overline{T},\ t = r,...,\overline{T}. \tag{5.2}$$

By iterated expectations, we can recover $\pi^r(t)$, the overall ATT for cohort r in post-treatment period t, by averaging (5.2) across the distribution of X_i. If the never-treated group can be considered a "good control" for each treated group conditional on X_i, we may use a “conditional staggered parallel trends assumption” (CSPTA), given as $\Delta^r_s(\infty, X_i) - \Delta^r_T(\infty, X_i) = 0$ for all $s = 1, ..., \overline{T}$ and $r \in \mathcal{R}$, where

$$\Delta^r_s(\infty, X_i) = \mathbb{E}[Y^\infty_i(s)|G^r_i = 1, D_{i,s} = 1, X_i] - \mathbb{E}[Y^\infty_i(s)|G^\infty_i = 1, D_{i,s} = 1, X_i].$$

This condition requires that, conditional on X_i, the development of the baseline potential outcome between time period t and the base period T in each treatment cohort matches the corresponding development in the never-treated group in the absence of treatment. Following (Wooldridge 2021, Section 7), we can allow for heterogeneity due to staggered treatment assignment as well as differences in observed characteristics by assuming that the observed data is generated as

$$\begin{aligned} Y_i = c + \kappa' X_i + \sum_{r \in \mathcal{R} \setminus \{\infty\}} (\alpha_r G^r_i + \zeta'_r X_i G^r_i) + \sum_{\substack{l=1 \\ l \neq T}}^{\overline{T}} (\gamma_l D_{i,l} + \xi'_l X_i D_{i,l}) \\ + \sum_{m=T+1}^{\overline{T}} \sum_{\substack{k=1 \\ k \neq T}}^{m-1} (\tau_{mk} G^m_i D_{i,k} + \upsilon'_{mk} \dot{X}^r_i G^m_i D_{i,k}) + \sum_{r=T+1}^{\overline{T}} \sum_{s=r}^{\overline{T}} (\tau_{rs} G^r_i D_{i,s} + \nu'_{rs} \dot{X}^r_i G^r_i D_{i,s}) + u_i, \end{aligned} \tag{5.3}$$

where $\dot{X}^r_i = X_i - \mathbb{E}[X_i|G^r_i = 1]$. Clearly, model (5.3) implicitly imposes that $\pi^r_X(t)$ is a linear function of X_i. However, the vector X_i may contain polynomial functions of the observed covariates. Simple algebra shows that $\tau_{rs} + \nu'_{rs} \dot{X}^r_i = \Delta^r_s(X_i) - \Delta^r_T(X_i)$, where

$$\Delta^r_s = \mathbb{E}[Y_i|G^r_i = 1, D_{i,s} = 1, X_i] - \mathbb{E}[Y_i|G^\infty_i = 1, D_{i,s} = 1, X_i].$$

Combining (5.3) with (5.1), the CSPTA implies that $\tau_{rs} + \nu'_{rs} \dot{X}^r_i = \pi^r_X(s)$. Due to the centering of X_i around its cohort mean, averaging across the distribution of X_i conditional on $G^r_i = 1$ shows that τ_{rs} can indeed be interpreted as the ATT for the treatment cohort r observed in period s. The model also includes placebo treatment effects τ_{mk} for individuals in cohort m observed in period $k < m$, i.e. before their actual treatment. If (5.1) and (5.3) hold, the CSPTA implies that $\tau_{mk} + \nu'_{mk} \dot{X}^m_i = 0$. We therefore avoid any “contamination” by treatment effects at time $m' > m$, which, as noted by Sun & Abraham (2021), can lead to a rejection of the CSPTA in the pre-treatment periods even in cases where it actually holds. Since asymptotic normality of the OLS estimator applied to model (5.3) holds under small modifications of the notation in Assumption 3.1, we can directly apply our equivalence tests. For instance, assuming that (5.3) and (5.1) hold, one can find evidence in favor of the unconditional staggered PTA by testing the null hypothesis that the maximum component of the vector $\tau_{placebo}$, defined as the vector collecting all τ_{mk} for $m = T+1, ..., \overline{T}$ and $k = 1, ..., m-1, k \neq T$, exceeds a certain threshold. If interest lies in the CSPTA, one can apply the same testing strategy on $\tau_{placebo}(x)$ with components $\tau_{mk} + \nu'_{mk}x$, where x denotes a specific outcome of X_i. In many applications, X_i is a single discrete variable, e.g. $X_i = 1$ for high-skilled workers and zero otherwise. One can then alternatively check the CSPTA by estimating (5.3) for both subgroups (omitting any interaction terms involving X_i) and applying our tests to the corresponding estimates of $\tau_{placebo}$. In principle, our methodology can also be used to test for negligible treatment effect heterogeneity by applying the same testing strategy to $\tau_{staggered}$, defined as the vector collecting all τ_{rs} for $r = T+1, ..., \overline{T}$, $s = r, ..., \overline{T}$.

The model in (5.3) can be flexibly adjusted to the problem at hand. For instance, one may be willing to exclude a subset of the placebo treatment effects from the model in order to allow for some pooling across cohorts and time. As noted by Wooldridge (2021), in this case, the pooled OLS estimator of τ_{rs} is an averaged “rolling DiD” where, on top of the never-treated group and the base period, any cohort and period that corresponds to an omitted placebo treatment effect is used as a control. In this case, the CSPTA needs to be adjusted accordingly (e.g. as in Roth et al. 2022), as parallel trends need to be plausible between multiple groups and periods. Finally, notice that in practice $\mathbb{E}[X_i|G_i^r = 1]$ needs to be replaced by the sample average of X_i in cohort r. As suggested in Wooldridge (2021), one should adjust the standard errors to account for the additional sampling variation.

6 Simulations

In order to investigate the small sample properties of our tests, we conduct a simulation study in **R**. For that, we create a data set of repeated cross sections, where the number of pre-treatment periods is $T \in \{2, 4, 8, 12\}$ and the number of individuals observed in each period n_t is either 100 or 1000. We set $\mathbb{P}(G_i = 1) = 0.5$ and $\mathbb{P}(D_l = 1) = 1/(T+1)$ in all simulations. Consequently, the treatment and the control group consist each of roughly half of the individuals and about the same number of individuals is observed in each period. We set the group dummy $\alpha = 2$ and draw the time dummies γ_l and the model error u_i independently from a standard normal distribution. Finally, we include an observed covariate X_i which is independently drawn from a normal distribution with mean and standard deviation 1.

In an initial step, we investigate the level of the proposed tests. To do so, we set the level of significance $\alpha = 5\%$ choose the threshold for all hypotheses as 1. We then choose the parameters β_l in the pre-treatment periods such that we are on the "boundary" of the hypotheses, that is $\beta_l = 1$ for some $l \in \{1, \ldots, T-1\}$ or $\beta_l = 1$ for all $l = 1, ..., T-1$. Moreover, we also investigate the power of the test procedures by choosing $\beta_l \in \{0.8, 0.9\}$ for all $l = 1, ..., T-1$. The bootstrap based test (3.11) for (3.1) is computed using 500 bootstrap draws. The results for all tests based on 20000 simulations are presented in Tables 2, 3 and 4.

In the following scenarios, we choose the level of significance $\alpha = 5\%$ and compute δ^*_{IU} and δ^*_{Boot} as the smallest equivalence thresholds for the intersection-union and the bootstrap tests such that the null hypothesis in (3.1) can still be rejected (i.e. for which equivalence of pre-trends can be concluded). Similarly, we compute the smallest equivalence thresholds τ^* and ζ^* for the corresponding null hypotheses in (3.2) and (3.3) using the tests in (3.16) and (3.21), respectively. The reported numbers correspond to the average over $M = 2500$ simulations and can be used to assess at what value of the equivalence threshold a particular test can be expected to reject the null hypothesis. Finally, we report the usual 95% confidence interval $\mathrm{CI}^{\hat{\beta}_{T+1}}$ and the number of simulations in which each β_l for $l = 1, ..., T-1$ was found to be statistically insignificant. In all our scenarios we set $\beta_{T+1} = 0$ so that the treatment has no effect. We then investigate how violations of the PTA affect the chance of falsely detecting a treatment effect and how these violations affect the smallest equivalence thresholds for which equivalence can be concluded.

Table 5 shows the results under the PTA. We further simulate scenarios in which the PTA is violated due to the presence of unobserved covariates that affect the treatment group but not the control group. Our first setup is Example 4.1 augmented by an additionally observed covariate X_i. The unobserved variable is modeled as $Z_i = G_i \times D_{i,T} \times V_i$, where

V_i denotes a random draw from a normal distribution with mean $\mu \in \{\frac{1}{4}, \frac{1}{2}\}$ and variance 1. The results are given in Table 6 and 7. The second setup includes a linear time trend as in Example 4.2, i.e. $Z_i = \psi \times t \times D_{i,t} \times G_i$ with $\psi \in \{0.025, 0.05\}$. The results are presented in Tables 8 and 9.

6.1 Simulation results – Discussion

Table 2 shows that the test in (3.16) approximately keeps the desired level for every T even in small samples. The test in (3.21) appears to be slightly over-rejecting when $n_t = 100$ but keeps its nominal level in larger samples. Notice that in Table 3 the tests in (3.16) and (3.21) rightfully reject the null hypothesis in an increasing number of cases as n_t and T increase. This makes sense, since an increase in $T > 2$ means that the average and the root mean square value of $\beta^{(T-1)}$ are further away from the boundary of the null, resulting in an increase in statistical power. Regarding the two tests in (3.9) and (3.11), Tables 2 and 3 illustrate that they maintain their nominal level for $T = 2$. When only one parameter is at the boundary of the null hypothesis, i.e. $\beta_1 = 1$ while $\beta_2 = ... = \beta_{T-1} = 0$, both tests perform well in the sense that the empirical rejection frequency is close to the nominal level for sufficiently large n. In contrast, if $\beta_l = 1$ for all pre-treatment periods $l = 1, ..., T-1$, both tests become conservative for larger values of T. This phenomenon appears to be much more pronounced for the test based on the intersection-union principle, for which it is well-documented (Berger & Hsu 1996). For instance, the empirical level of the intersection-union test is more than 6 times smaller than the corresponding level of the bootstrap based test for $T = 8$. As shown in Table 4, this has important consequences for the power of both tests. As can easily be seen, our bootstrap based test procedure outperforms the intersection-union-based test for $T > 2$. On the other hand, the intersection-union based test may still be attractive for practical applications at it is numerically much less demanding. As compared to the tests in (3.9) and (3.11), the power of our test in (3.21) is substantially larger, only surpassed by the power of the test in (3.16). All tests have in common that the power decreases with T. This is true even for the test in (3.16), which is the uniformly most powerful (asymptotic) test for this null hypothesis for any T. Thus, concluding equivalence of pre-trends becomes more demanding with an increase in the number of pre-treatment periods. This makes intuitive sense in the DiD setup, where equivalence of pre-trends in a larger number of periods is often regarded as stronger evidence for the plausibility of the PTA.

The fact that the intersection-union test for (3.1) becomes very conservative may explain why δ^*_{IU} is increasing in T for all sample sizes and in all simulation setups. One of the reasons for this behavior is that due to its construction, the value of δ^*_{IU} is largely determined by the maximal variation in the components of $(\hat{\beta}_1, ..., \hat{\beta}_{T-1})$. For $T = 2$, we see that $\zeta^* > \delta^*_{IU} \approx \delta^*_{Boot} = \tau^*$. This may be explained by the fact that for $T = 2$ the tests in (3.1) and (3.16) coincide while (asymptotically) being the uniformly most powerful test. The bootstrap based test performs very well, as anticipated from the high power shown in Table 4. For $T > 2$, we roughly observe that $\tau^* < \zeta^* \leq \delta^*_{Boot} < \delta^*_{IU}$, which can be explained by the observation $\bar{\beta}^{(T-1)} \leq \beta_{RMS} \leq \|\beta^{(T-1)}\|_\infty$. In all cases, δ^*_{Boot} is substantially smaller than δ^*_{IU}, which may be explained by the higher power of the bootstrap based test. Further notice that even when the PTA holds, the practice of rejecting the DiD framework when $\hat{\beta}_l$ is statistically insignificant for at least one $l \in \{1, ..., T-1\}$ is clearly inefficient as is shown by the first row of Table 5, as an increase in available pre-treatment periods increases the chance of incorrectly rejecting the DiD framework under the PTA. Thus, rather than rejecting a DiD analysis in an application due to a significant pre-treatment parameter estimate, it may be more sensible to use an equivalence test based procedure. A similar observation can be made in the presence of a linear time trend as shown in Tables 8 and 9. Here, even when the empirical coverage rate of the usual confidence interval is only slightly lower than the nominal level, the DiD framework is rejected in a large number of cases.

When the PTA is violated due to a small temporary shock as in Table 6, the usual practice of adopting the PTA when no significant differences in pre-trends could be found can lead to a false discovery of a non-zero treatment effect in a substantial number of cases, in particular when the sample size is small. If the temporal shock is larger as in Table 7, a non-existing treatment effect will be found to be significantly different from zero in almost all cases. All our test procedures require an unrealistically large equivalence threshold in order to be able to conclude equivalence of pre-trends. In particular, any equivalence threshold for which equivalence could be concluded would have to be larger than the estimated treatment effect, therefore casting serious doubt on the validity of the latter. We further notice that δ^*_{IU} and δ^*_{Boot} increase in T, whereas τ^* and ζ^* remain stable or slightly decrease in T, as variation in $\hat{\beta}^{(T-1)}$ is "smoothed out" with more pre-treatment periods available.

When the PTA is violated due to a permanent linear time trend that affects only the treatment group, the bias of $\hat{\beta}_{T+1}$ corresponds to the slope of the trend. Again, when the sample is large (an thus when the width of $\mathrm{CI}^{\hat{\beta}_{T+1}}$ is small), $\mathrm{CI}^{\hat{\beta}_{T+1}}$ contains the true ATT in less than 95% of the cases. If the slope of the time trend is small, the coverage of $\mathrm{CI}^{\hat{\beta}_{T+1}}$ is however close to its nominal level, as is shown in Table 8. As expected from Example 4.2, the coverage gets worse with a steeper slope of the time trends (Table 9). As before, in order to be able to conclude equivalence of pre-trends, the equivalence thresholds would have to be chosen larger than the estimated treatment effect, thus suggesting that the estimated ATT may contain bias due to insufficient support for the PTA. Moreover, our methodology can be useful in identifying the presence of a linear time trend, as τ^* and ζ^* tend do decrease with T under the PTA or when the violation of the PTA is only temporary, whereas under the presence of a linear trend, they increase with T (as is expected by Example 4.2).

7 Empirical illustration

In this section, we illustrate our approach by re-considering the influential Difference-in-Differences analysis in Di Tella & Schargrodsky (2004). They use a shock to the allocation of police forces as a consequence of a terrorist attack on a Jewish institution as a natural experiment to study the the effect of police on crime. We choose this paper as it provides an excellent opportunity for a comparison between our methodology and the current standard of testing for violations of parallel trends. The original authors conduct the usual Granger-causality test in (2.5) and find no evidence for violations of the PTA. However, Donohue et al. (2013) point out several shortcomings of the original paper (e.g. spillover effects from the treated to the untreated group). In particular, they find that the PTA is not plausible if the pre-treatment data is inspected on a more granular level, thus casting doubt on the validity of the estimated treatment effects. While the traditional test failed to detect evidence *against* the PTA, we will apply our test procedures to analyze how much evidence *in favor* of the PTA can be extracted from the original specification in Di Tella & Schargrodsky (2004).

The data consists of monthly averages of the number of car thefts between April and December 1994 in each out of 876 Buenos Aires city blocks out of which 37 blocks hosted Jewish institutions and thus received additional protection after the attack. The main specification in Di Tella & Schargrodsky (2004) is given by $Y_{it} = \alpha_i + \gamma_t + \beta D_{it}$, where Y_{it} denotes the number of car thefts in block i and month t and D_{it} is a dummy variable taking the value 1 if block i is treated in period t. Finally, α_i and γ_t are block- and time-specific fixed effects. By using this specification, the pre- and post-treatment periods are pooled together so that the estimated treatment effect compares the post-treatment difference in car thefts between treated and non-treated blocks to the corresponding pre-treatment difference. To analyze group mean differences in the pre-treatment periods, we adapt (2.4) by pooling the post-treatment periods in two different specifications. First, as in the original paper, we include block-specific effects and cluster on the block level. Secondly, we replace the block-specific dummies by a single group dummy and compute heteroskedasticity-robust standard errors. Finally, we compute δ^*_{IU}, δ^*_{Boot}, τ^* and ζ^* based on one, two and three pre-treatment periods, corresponding to June, May and June and April–June. Since the data set is a panel that is ordered by time (as are most panel data sets in practice), the implementation of our test in (3.21) needs to be adjusted slightly: instead of choosing the first $\lfloor \lambda n \rfloor$ observations in the data set, we use the first $\lfloor \lambda n/(T+1) \rfloor$ observations in each time period to compute $\hat{\theta}(\lambda)$ for $\lambda \in \{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\}$. Moreover, notice that (3.11) and (3.21) do not require an estimator of the asymptotic variance. Thus, they are not affected by the choice of standard errors. The results are summarized in Table 1 below. Notice that for all tests the smallest equivalence threshold that still allows us to conclude equivalence of pre-trends are the largest when only the pre-treatment period June is used. If more pre-treatment periods are taken into account, the minimum upper bounds δ^*_{IU} and δ^*_{Boot} stay constant whereas the average and the root mean squared upper bounds τ^* and ζ^* become smaller. This hints towards a temporary shock to treatment or control in June which may bias the pooled estimates in Table 3 of Di Tella & Schargrodsky (2004). The latter are significant and range between -0.058 and -0.081. One important outcome

Estimates \ periods	June	May&June	April–June
δ^*_{IU} (clustered)	0.104	0.104	0.104
δ^*_{IU} (White)	0.161	0.161	0.161
δ^*_{Boot}	0.156	0.156	0.156
τ^* (clustered)	0.104	0.083	0.076
τ^* (White)	0.161	0.098	0.093
ζ^*	0.106	0.076	0.066

Table 1: Smallest equivalence thresholds such that the null hypotheses in (3.1), (3.2) and (3.3) can be rejected for varying numbers of pre-treatment periods.

of our equivalence test based analysis is that, even without the granular data inspection of Donohue et al. (2013), the equivalence thresholds have to be chosen unrealistically large in order to conclude equivalence of pre-trends. In fact, the smallest equivalence bounds for which the null hypotheses can be rejected are larger than the estimated effect size of police on crime. Therefore, it is questionable whether there is any effect at all, since the estimated effect may be an artifact of the violated PTA only.

8 Conclusion

We have derived four distinct procedures for testing equivalence of pre-trends in difference-in-differences estimation. Our tests capture the maximum, average and root mean square change in group mean differences relative to the base period and thus provide a measure of similarity between treatment and control. Contrary to the current practice, our tests require researchers to provide evidence in support of the parallel trends assumption. Our approach is based on the explicit specification of a threshold below which equivalence can be assumed. Alternatively, we propose to compare the estimated treatment effects with the smallest equivalence threshold for which equivalence can still be concluded for a given level of significance. Computationally, our tests are based on simple linear regressions. Therefore, they can easily be adapted to more complicated setups, including heterogeneous treatment effects and staggered treatment assignment. In a simulation study, we further show that our tests maintain their nominal level and exhibit high statistical power in sufficiently large samples. Moreover, we illustrate the performance of our tests under violations of the parallel trends assumption. Finally, we apply our methodology to the data provided by Di Tella & Schargrodsky (2004). Even without a granular inspection of the data as in Donohue et al. (2013), our methodology casts doubt on the estimated effects, as they may simply result from previously undetected differences in pre-trends.

Acknowledgements

We thank the editor Ivan Canay, an associate editor, and two anonymous referees for comments that greatly improved this paper.

References

Angrist, J. D. & Pischke, J.-S. (2008), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Berger, R. L. & Hsu, J. C. (1996), 'Bioequivalence trials, intersection-union tests and equivalence confidence sets', *Statistical Science* **11**(4), 283–319.

Bilinski, A. & Hatfield, L. A. (2020), 'Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions'.

Borusyak, K., Jaravel, X. & Spiess, J. (2021), 'Revisiting event study designs: Robust and efficient estimation'.

Callaway, B. & Sant'Anna, P. H. (2021), 'Difference-in-differences with multiple time periods', *Journal of Econometrics* **225**(2), 200–230.

de Chaisemartin, C. & D'Haultfoeuille, X. (2022), Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey, Working Paper 29691, National Bureau of Economic Research.

de Chaisemartin, C. & D'Haultfœuille, X. (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**(9), 2964–96.

Dette, H., Kokot, K. & Aue, A. (2020), 'Functional data analysis in the Banach space of continuous functions', *Annals of Statistics* **48**, 1168–1192.

Dette, H., Kokot, K. & Volgushev, S. (2020), 'Testing relevant hypotheses in functional time series via self-normalization', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**(3), 629–660.

Dette, H., Möllenhoff, K., Volgushev, S. & Bretz, F. (2018), 'Equivalence of regression curves', *Journal of the American Statistical Association* **113**, 711–729.

Dette, H. & Wied, D. (2014), 'Detecting relevant changes in time series models', *Journal of the Royal Statistical Society: Series B* **78**(2), 371 – 394.

Dette, H. & Wu, W. (2019), 'Detecting relevant changes in the mean of nonstationary processes – a mass excess approach', *Annals of Statistics* **47**, 3578–3608.

Di Tella, R. & Schargrodsky, E. (2004), 'Do police reduce crime? estimates using the allocation of police forces after a terrorist attack', *American Economic Review* **94**(1), 115–133.

Donohue, J. J., Ho, D. & Leahy, P. (2013), 'Do police reduce crime? a reexamination of a natural experiment', *Empirical Legal Analysis: Assessing the Performance of Legal Institutions* pp. 125–143.

Goodman-Bacon, A. (2021), 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics* **225**(2), 254–277.

Hartman, E. & Hidalgo, F. D. (2018), 'An equivalence approach to balance and placebo tests', *American Journal of Political Science* **62**(4), 1000–1013.

Heckman, J. J. & Smith, J. A. (1999), 'The pre-programme earnings dip and the determinants of participation in a social programme. implications for simple programme evaluation strategies', *The Economic Journal* **109**(457), 313–348.

Kahn-Lang, A. & Lang, K. (2020), 'The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications', *Journal of Business & Economic Statistics* **38**(3), 613–620.

Merlevède, F., Peligrad, M. & Utev, S. (2006), 'Recent advances in invariance principles for stationary sequences.', *Probability Surveys* **3**, 1–36.

Rambachan, A. & Roth, J. (2022), A more credible approach to parallel trends, Technical report.

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

Roth, J. (2022), 'Pretest with caution: Event-study estimates after testing for parallel trends', *American Economic Review: Insights* **4**(3), 305–22.

Roth, J., Sant'Anna, P. H. C., Bilinski, A. & Poe, J. (2022), 'What's trending in difference-in-differences? a synthesis of the recent econometrics literature'.

Sun, L. & Abraham, S. (2021), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* **225**(2), 175–199.

van der Vaart, A. & Wellner, J. A. (1996), *Weak convergence and empirical processes: With applications to statistics*, Springer-Verlag, New York.

Wellek, S. (2010), *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edn, CRC Press, Boca Raton, FL.

Wing, C., Simon, K. & Bello-Gomez, R. A. (2018), 'Designing difference in difference studies: Best practices for public health policy research', *Annual Review of Public Health* **39**(1), 453–469.

Wooldridge, J. M. (2021), 'Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators', *Available at SSRN 3906345* .

A Mathematical proofs

A.1 Properties of the test (3.8)

For sufficiently large sample sizes the quantile $f_\alpha := \mathcal{Q}_{\mathrm{N}_F(\delta,\hat{\Sigma}_{11}/n)}(\alpha)$ satisfies

$$\alpha = \mathbb{P}\big(\mid \mathrm{N}_F(\delta, \hat{\Sigma}_{11}/n) \mid \leq \mathcal{Q}_{\mathrm{N}_F(\delta,\hat{\Sigma}_{11}/n)}(\alpha)\big) = \Phi\Big(\frac{f_\alpha - \delta}{\Sigma_{11}}\Big) - \Phi\Big(\frac{-f_\alpha - \delta}{\Sigma_{11}}\Big) + \mathrm{O}\left(\frac{1}{\sqrt{n}}\right) \quad \text{(A.1)}$$

where Φ is the cdf of the standard normal distribution. Consequently, we obtain for the probability of rejection

$$\mathbb{P}_{\beta_1}(|\hat{\beta}_1| \leq f_\alpha) \approx \Phi\Big(\frac{f_\alpha - \beta_1}{\Sigma_{11}}\Big) - \Phi\Big(\frac{-f_\alpha - \beta_1}{\Sigma_{11}}\Big). \quad \text{(A.2)}$$

It is well known that the right-hand side of (A.2) (with the quantile f_α defined by (A.1)) is the power function of the uniformly most powerful unbiased test (see Example 1.1 in Romano (2005)).

A.2 Proof of Theorem 3.1

The proof follows essentially by the same arguments as given in Dette et al. (2018), and, for the sake of brevity, we only explain why this is the case. First note that a standard calculation (see also the discussion below in Section A.3, where a sequential version of the result is derived) shows that

$$\sqrt{n}(\hat{\theta} - \theta) = \Gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i u_i + o_{\mathbb{P}}(1) \ ,$$

where $\hat{\theta} = \hat{\Gamma}^{-1} \frac{1}{n} \sum_{i=1}^{n} W_i Y_i$ is the OLS of the parameter θ in model (2.4) from the observations $(W_1, Y_1), \ldots, (W_n, Y_n)$ and $\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} W_i W_i'$. Consequently, by the CLT $\sqrt{n}(\hat{\theta} - \theta)$ has an asymptotic normal distribution. Observing the definition of the vector $\beta^{(T-1)}$ as a sub-vector of θ, it follows from the continuous mapping theorem that $\sqrt{n}(\hat{\beta}^{(T-1)} - \beta^{(T-1)})$ has an asymptotic $(T-1)$-dimensional centred normal distribution as well. We denote the corresponding asymptotic covariance matrix by $\Sigma^{(T-1)} = (\sigma_{ij})_{i,j=1,\ldots T-1}$. Now we interpret all vectors as stochastic processes on the set $\mathcal{X} = \{1, \ldots, T-1\}$ and rewrite the weak convergence of the vector $\hat{\beta}^{(T-1)} = (\hat{\beta}_1, \ldots, \hat{\beta}_{T-1})'$ as

$$\{\sqrt{n}(\hat{\beta}_x - \beta_x)\}_{x \in \mathcal{X}} \rightsquigarrow \{\mathbb{G}(x)\}_{x \in \mathcal{X}}, \quad \text{(A.3)}$$

where $\{\mathbb{G}(x)\}_{x\in\mathcal{X}}$ is a centered Gaussian process on $\mathcal{X} = \{1,\dots,T-1\}$ with covariance structure $\mathrm{Cov}(\mathbb{G}(x),\mathbb{G}(y)) = \sigma_{xy}$ $(x, y \in \mathcal{X})$. Note that (A.3) is the analog of equation (A.7) in Dette et al. (2018), and it follows by exactly the same arguments as stated in this paper that

$$\sqrt{n}\big(\|\hat{\beta}^{(T-1)}\|_\infty - \|\beta^{(T-1)}\|_\infty\big) \to \max\big\{\max_{x\in\mathcal{E}^+}\mathbb{G}(x), \max_{x\in\mathcal{E}^-} -\mathbb{G}(x)\big\}\ , \tag{A.4}$$

provided that $\|\hat{\beta}^{(T-1)}\|_\infty > 0$, where the sets $\mathcal{E}^+$ and $\mathcal{E}^-$ are defined by

$$\begin{aligned}\mathcal{E}^+ &= \{\ell = 1,\dots,T-1:\ \beta_\ell = \|\beta^{(T-1)}\|_\infty\}\ ,\\ \mathcal{E}^- &= \{\ell = 1,\dots,T-1:\ \beta_\ell = -\|\beta^{(T-1)}\|_\infty\}\ ,\end{aligned}$$

respectively. Note that $\mathcal{E}^- \cup \mathcal{E}^+ = \mathcal{E}$, where $\mathcal{E}$ is defined in (3.13), and that (A.3) is the analog of Theorem 3 in Dette et al. (2018). Moreover, if $\hat{\beta}^{(T-1),*} = (\hat{\beta}^*_1,\dots,\hat{\beta}^*_{T-1})'$ denotes the estimate from the bootstrap sample, we obtain an analog of the weak convergence in (A.3), that is

$$\{\sqrt{n}(\hat{\beta}^*_x - \hat{\hat{\beta}}_x)\}_{x\in\mathcal{X}} \rightsquigarrow \{\mathbb{G}(x)\}_{x\in\mathcal{X}} \tag{A.5}$$

conditional on the sample $(W_1, Y_1),\dots,(W_n, Y_n)$. Note that this statement corresponds to the statement (A.25) in Dette et al. (2018). Now the statements (A.7) and (A.25) and their Theorem 3 are the main ingredients for the proof of Theorem 5 in Dette et al. (2018). In the present context these statements can be replaced by (A.3), (A.5) and (A.4), respectively, and a careful inspection of the arguments given in Dette et al. (2018) shows that Theorem 3.1 holds (the arguments even simplify substantially as in our case the index set $\mathcal{X}$ of the processes is finite).

A.3 Proof of Theorem 3.2

Recall that $\hat{\theta}(\lambda)$ is the OLS for the parameter θ in model (2.4) from the observations $(W_1, Y_1),\dots,(W_{\lfloor n\lambda\rfloor}, Y_{\lfloor n\lambda\rfloor})$, that is

$$\hat{\theta}(\lambda) \ = \ \hat{\Gamma}^{-1}_{\lfloor n\lambda\rfloor}\frac{1}{\lfloor n\lambda\rfloor}\sum_{i=1}^{\lfloor n\lambda\rfloor} W_iY_i = \theta + \hat{\Gamma}^{-1}_{\lfloor n\lambda\rfloor}\frac{1}{\lfloor n\lambda\rfloor}\sum_{i=1}^{\lfloor n\lambda\rfloor} W_iu_i,$$

where the matrix Γ_k is defined by

$$\hat{\Gamma}_k = \frac{1}{k}\sum_{i=1}^{k} W_i W_i'.$$

As

$$\sup_{\lambda\in[\varepsilon,1]} \|\hat{\Gamma}_{\lfloor n\lambda\rfloor} - \Gamma\| = o_{\mathbb{P}}(1)$$

and the matrix Γ is non-singular, it follows that

$$\sqrt{n}(\hat{\theta}(\lambda) - \theta) = \Gamma^{-1}\frac{\sqrt{n}}{\lfloor n\lambda\rfloor}\sum_{i=1}^{\lfloor n\lambda\rfloor} W_i u_i + o_{\mathbb{P}}(1)$$

uniformly with respect to $\lambda \in [\varepsilon, 1]$. Consequently, we obtain from the Cramer-Wold device and Theorem 2.12.1 in van der Vaart & Wellner (1996) that

$$\big\{\sqrt{n}(\hat{\theta}(\lambda) - \theta)\big\}_{\lambda\in[\varepsilon,1]} \rightsquigarrow \Big\{\frac{\eta\Gamma^{-1/2}}{\lambda}\vec{\mathbb{B}}(\lambda)\Big\}_{\lambda\in[\varepsilon,1]} \tag{A.6}$$

where $\vec{\mathbb{B}}$ is a $2T+2$-dimensional vector of independent Brownian motions, $\eta = \mathrm{Var}(u_i)$ and the symbol $\rightsquigarrow$ means weak convergence in the space $(\ell^\infty[\varepsilon,1])^{2T+2}$ of all $(2T+2)$-dimensional bounded functions on the interval $[\varepsilon, 1]$. As the projections of θ on its coordinates are continuous mappings, the weak convergence (A.6) and the continuous mapping theorem imply

$$\big\{\sqrt{n}(\hat{\beta}^{(T-1)}(\lambda) - \beta^{(T-1)})\big\}_{\lambda\in[\varepsilon,1]} \quad\rightsquigarrow\quad \big\{\frac{1}{\lambda}D\vec{\mathbb{B}}(\lambda)\big\}_{\lambda\in[\varepsilon,1]}, \tag{A.7}$$

where D is a $(T-1)\times(2T+2)$ matrix of full rank. In the case $\beta^{(T-1)} = 0$ the result in Theorem 3.2 now follows directly from the continuous mapping theorem. On the other hand, if $\beta^{(T-1)} \neq 0$, it follows that

$$\begin{aligned} H_n(\lambda) &= \sqrt{n}\big(\|\hat{\beta}^{(T-1)}(\lambda)\|^2 - \|\beta^{(T-1)}\|^2\big) \\ &= \sqrt{n}\{\|\hat{\beta}^{(T-1)}(\lambda) - \beta^{(T-1)}\|^2 + 2(\hat{\beta}^{(T-1)}(\lambda) - \beta^{(T-1)})'\beta^{(T-1)} \\ &= 2\sqrt{n}(\hat{\beta}^{(T-1)}(\lambda) - \beta^{(T-1)})'\beta^{(T-1)} + o_{\mathbb{P}}(1) \end{aligned}$$

uniformly with respect to $\lambda \in [\varepsilon, 1]$, and a further application of the continuous mapping theorem yields

$$\big\{H_n(\lambda)\big\}_{\lambda\in[\varepsilon,1]} \rightsquigarrow \Big\{2(\beta^{(T-1)})'D\frac{\vec{\mathbb{B}}(\lambda)}{\lambda}\Big\}_{\lambda\in[\varepsilon,1]}$$

in $\ell^\infty([\varepsilon, 1])$. It is easy to see that for $(\beta^{(T-1)}) \neq 0$ the process on the right-hand side equals in distribution

$$\Big\{\Delta(\beta^{(T-1)})\frac{\mathbb{B}_1(\lambda)}{\lambda}\Big\}_{\lambda\in[\varepsilon,1]}$$

where $\mathbb{B}_1$ is a one-dimensional Brownian motion and

$$\Delta(\beta^{(T-1)}) = 4(\beta^{(T-1)})' DD' \beta^{(T-1)} \tag{A.8}$$

is a positive constant. Recalling the definition of the statistic $\hat{M}_n$ in (3.17) and a further application of the continuous mapping theorem shows that

$$\begin{aligned}
\hat{M}_n &= \frac{\hat{\beta}^2_{RMS}(1) - \beta^2_{RMS}}{\hat{V}_n} \\
&= \frac{\|\hat{\beta}^{(T-1)}(1)\|^2 - \|\beta^{(T-1)}\|^2}{\big(\int_\varepsilon^1 (\|\hat{\beta}^{(T-1)}(\lambda)\|^2 - \|\hat{\beta}^{(T-1)}(1)\|^2)^2 \nu(d\lambda)\big)^{1/2}} \\
&= \frac{H_n(1)}{\big(\int_\varepsilon^1 (H_n(\lambda) - H_n(1))^2 \nu(d\lambda)\big)^{1/2}} \\
&\overset{d}{\to} \mathbb{W} = \frac{\mathbb{B}_1(1)}{\big(\int_\varepsilon^1 (\mathbb{B}_1(\lambda)/\lambda - \mathbb{B}_1(1))^2 \nu(d\lambda)\big)^{1/2}},
\end{aligned}$$

which proves the assertion.

A.4 Proof of Theorem 3.3

Observing the definition of $\hat{M}_T$ in (3.17) we obtain

$$\mathbb{P}_{\beta^{(T-1)}}\Big(\hat{\beta}^2_{RMS} < \zeta^2 + \mathcal{Q}_{\mathbb{W}}(\alpha)\hat{V}_n\Big) = \mathbb{P}_{\beta^{(T-1)}}\Big(\hat{M}_n < \frac{\zeta^2 - \beta^2_{RMS}}{\hat{V}_n} + Q_{\mathbb{W}}(\alpha)\Big).$$

It follows from the proof of Theorem 3.2 that $\hat{V}_n = O_{\mathbb{P}}(1/\sqrt{n})$. Consequently, if $\beta^2_{RMS} > 0$, the assertion follows by a simple calculation considering the three cases separately. On the other hand, if $\beta_{RMS} = 0$, the proof of Theorem 3.2 also shows that $\|\hat{\beta}^{(T-1)}(1)\|^2 = O_{\mathbb{P}}(\frac{1}{n})$ and the assertion follows from the weak convergence (3.20) in Theorem 3.2.

B Simulation results

	$n_t = 100$				$n_t = 1000$			
Test	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
(3.9)	0.0502	0.0049	0.0005	0.0004	0.0502	0.0051	0.0010	0.0000
(3.11)	0.0512	0.0154	0.0071	0.0046	0.0527	0.0132	0.0066	0.0046
(3.16)	0.0502	0.0474	0.0523	0.0510	0.0502	0.0481	0.0507	0.0491
(3.21)	0.0983	0.0815	0.0725	0.0764	0.0599	0.0556	0.0595	0.0555

Table 2: Rejection frequencies for $\beta_t = 1$, $t = 1, ..., T - 1$ with equivalence threshold 1 at nominal level of significance $\alpha = 5\%$.

	$n_t = 100$				$n_t = 1000$			
Test	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
(3.9)	0.0483	0.0389	0.0239	0.0197	0.0508	0.0493	0.0503	0.0512
(3.11)	0.0546	0.0783	0.0944	0.1052	0.0512	0.0527	0.0506	0.0505
(3.16)	0.0483	0.8902	0.9896	0.9965	0.0508	1.0000	1.0000	1.0000
(3.21)	0.0964	0.5739	0.8182	0.8610	0.0585	0.9979	1.0000	1.0000

Table 3: Rejection frequencies for $\beta_1 = 1$ and $\beta_l = 0$, $l = 2, ..., T - 1$ with equivalence threshold 1 at nominal level of significance $\alpha = 5\%$.

	$\beta_l = 0.8,\, l = 1, ..., T-1$				$\beta_l = 0.9,\, l = 1, ..., T-1$			
Test	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
(3.9)	1.0000	0.5915	0.3609	0.2750	1.0000	0.1525	0.0410	0.0215
(3.11)	1.0000	0.6331	0.5129	0.4740	1.0000	0.2033	0.1114	0.0982
(3.16)	1.0000	1.0000	0.9983	0.9934	1.0000	1.0000	0.9580	0.8676
(3.21)	1.0000	0.9768	0.9012	0.8617	1.0000	0.9043	0.6446	0.5418

Table 4: Rejection frequencies for $n_t = 1000$ with equivalence threshold 1 at nominal level of significance $\alpha = 5\%$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.9476	0.8684	0.7696	0.6912	0.9492	0.8584	0.7604	0.7116
$\hat{\beta}_{T+1}$	0.0061	0.0015	-0.0005	-0.0020	0.0030	0.0020	0.0000	-0.0018
$\mathrm{CI}^{\hat{\beta}_{T+1}}$	0.9548	0.9444	0.9524	0.9435	0.9484	0.9484	0.9504	0.9540
δ^*_{IU}	0.6360	0.8216	0.9148	0.9536	0.2002	0.2581	0.2888	0.3017
δ^*_{Boot}	0.6460	0.6644	0.6455	0.6171	0.2022	0.2137	0.2099	0.1998
τ^*	0.6360	0.5207	0.4782	0.4675	0.2002	0.1651	0.1515	0.1478
ζ^*	0.7104	0.6958	0.6923	0.7016	0.2196	0.2141	0.2099	0.2091

Table 5: Estimation and test performance under the PTA at nominal level of significance $\alpha = 5\%$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.8636	0.7028	0.5426	0.4461	0.2756	0.0912	0.0372	0.0192
$\hat{\beta}_{T+1}$	-0.2563	-0.2592	-0.2648	-0.2509	-0.2520	-0.2512	-0.2528	-0.2537
$\mathrm{CI}^{\hat{\beta}_{T+1}}$	0.8560	0.8347	0.8244	0.8398	0.2640	0.2460	0.2400	0.2344
δ^*_{IU}	0.8016	0.9900	1.0928	1.1242	0.4155	0.4671	0.5028	0.5171
δ^*_{Boot}	0.7885	0.8412	0.8330	0.8331	0.4102	0.4371	0.4570	0.4599
τ^*	0.8016	0.7153	0.6963	0.6693	0.4155	0.3895	0.3856	0.3843
ζ^*	0.9279	0.8334	0.8050	0.8211	0.4216	0.4032	0.3975	0.3938

Table 6: Estimation and test performance under violation of the PTA due to a temporary group-specific shock ($Z_{ist} = G_i \times D_T \times V_i$ with $V_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(\frac{1}{4}, 1)$) at nominal level of significance $\alpha = 5\%$ with $\beta_{T+1} = 0$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.6036	0.4024	0.2565	0.1797	0.0012	0.0000	0.0000	0.0000
$\hat{\beta}_{T+1}$	-0.5028	-0.4959	-0.5023	-0.4976	-0.4983	-0.4970	-0.4991	-0.50004
$\mathrm{CI}^{\hat{\beta}_{T+1}}$	0.6176	0.6018	0.5810	0.5910	0.0008	0.0004	0.0000	0.0004
δ^*_{IU}	1.0299	1.1943	1.2953	1.3503	0.6673	0.7161	0.7495	0.7638
δ^*_{Boot}	1.0040	1.0862	1.1144	1.1312	0.6639	0.6843	0.7011	0.7126
τ^*	1.0299	0.9424	0.9163	0.9181	0.6673	0.6391	0.6329	0.6303
ζ^*	1.0357	1.0098	0.9984	1.0064	0.6869	0.6692	0.6645	0.6614

Table 7: Estimation and test performance under violation of the PTA due to a temporary group-specific shock ($Z_{ist} = G_i \times D_T \times V_i$ with $V_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(\frac{1}{2}, 1)$) at nominal level of significance $\alpha = 5\%$ with $\beta_{T+1} = 0$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.9460	0.8635	0.7015	0.5494	0.9376	0.7848	0.2884	0.0236
$\hat{\beta}_{T+1}$	0.0269	0.0267	0.0207	0.0284	0.0235	0.0267	0.0258	0.0234
$\text{CI}^{\hat{\beta}_{T+1}}$	0.9528	0.9484	0.9520	0.9476	0.9372	0.9392	0.9388	0.9388
δ^*_{IU}	0.6332	0.8227	0.9382	1.0209	0.2050	0.2764	0.3652	0.4606
δ^*_{Boot}	0.6348	0.6763	0.6875	0.6875	0.2049	0.2404	0.3242	0.4354
τ^*	0.6332	0.5261	0.5029	0.5177	0.2050	0.1815	0.2115	0.2591
ζ^*	0.7178	0.7136	0.7141	0.7221	0.2197	0.2229	0.2496	0.2899

Table 8: Estimation and test performance under violation of the PTA due to a time trend with slope 0.025 ($Z_i = 0.025 \times t \times D_{i,t} \times G_i$) at nominal level of significance $\alpha = 5\%$ with $\beta_{T+1} = 0$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.9458	0.8392	0.5501	0.2353	0.9130	0.5271	0.0061	0.0000
$\hat{\beta}_{T+1}$	0.0544	0.0593	0.0547	0.0457	0.0514	0.0495	0.0530	0.0504
$\text{CI}^{\hat{\beta}_{T+1}}$	0.9420	0.9436	0.9445	0.9495	0.9195	0.9120	0.9135	0.9096
δ^*_{IU}	0.6395	0.8437	1.0144	1.1961	0.2144	0.3233	0.5150	0.7158
δ^*_{Boot}	0.6528	0.7066	0.8251	1.0038	0.2166	0.2979	0.5029	0.7039
τ^*	0.6395	0.5412	0.5612	0.6478	0.2144	0.2205	0.3098	0.4084
ζ^*	0.7212	0.7225	0.7484	0.8030	0.2271	0.2505	0.3466	0.4575

Table 9: Estimation and test performance under violation of the PTA due to a time trend with slope 0.05 ($Z_i = 0.05 \times t \times D_{i,t} \times G_i$) at nominal level of significance $\alpha = 5\%$ with $\beta_{T+1} = 0$.