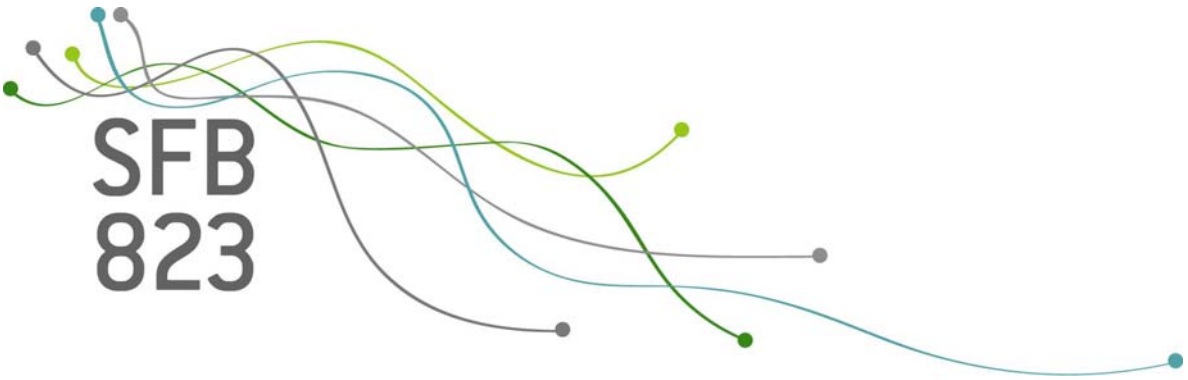
SFB 823

Difference-in-differences estimation under non-parallel trends

Holger Dette, Martin Schumann

Nr. 22/2020

Discussion Paper



DIFFERENCE-IN-DIFFERENCES ESTIMATION UNDER NON-PARALLEL TRENDS

HOLGER DETTE AND MARTIN SCHUMANN

Abstract. Classic difference-in-differences estimation relies on the validity of the "parallel trends assumption" (PTA), which ensures that the evolution of the variable of interest in the control group can be used to determine its counterfactual development in the treatment group in the absence of treatment. The plausibility of the PTA is usually assessed by a test of the null hypothesis that the difference between the means of both groups is constant over time before the treatment. However, this procedure is problematic as failure to reject the null hypothesis does not imply the absence of differences in time trends between both groups due to low power to detect economically relevant differences. We provide three tests of equivalence leading to a "common range" (CR) condition that replaces the PTA and which naturally reflects differences between treatment and control. We combine the CR with standard confidence intervals to capture both design and sampling uncertainty in the data and show that the combined confidence intervals yield more reliable inference when the PTA is violated.

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 DiD estimators yields sensible point estimates of the true effect of the treatment is known as the "parallel trends" or "parallel paths" assumption, which states that in the absence of treatment on average both groups would have experienced the same temporal trends in the outcome variable. 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 in favor 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 and Lang, 2019). 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 (2020) 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. This is in line with Ioannidis, Stanley, and Doucouliagos (2017), who argue that lack of statistical power

Compiled: 11th August, 2020 at 3:34pm.

is not only making it difficult to find evidence for relationships that exist in the population but may also lead to an increase in the rates of false positives, where statistical evidence is reported for non-existent relationships. In summary, while finding evidence for differences in trends in pre-treatment periods is a strong argument against exactly parallel trends, failure to detect different trends is an insufficient justification for the use of the standard DiD estimation procedure.

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".[1] We provide three distinct types of equivalence that impose bounds on the maximum, the average and the "least squares" change over time in the group mean difference between treatment and control in the pre-treatment periods. These bounds thus form a "common range" for the changes in group mean differences between treatment and control over time. We then proceed by relaxing the usual parallel trends assumption by assuming that in the absence of treatment differences in trends between treatment and control in the post-treatment period are bounded by the common range. This naturally implies that the average treatment effect on the treated (ATT) is no longer point-identified but lies within a range that naturally reflects the (dis-) similarities between both groups. We finally combine the usual confidence interval around the treatment effect estimate (measuring sampling uncertainty) with the common range (measuring design uncertainty) to obtain simple intervals that capture both sources of uncertainty. This procedure has several advantages over the current standard pre-test. First, it 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. Second, due to our reversal of the standard null hypothesis, we are able to quantify the extend to which the two groups are comparable in the pre-treatment periods, as differences are bounded by the common range with probability of at least $1-\alpha$ for a given level of significance α. Third, the width of the common range constitutes a measure of the plausibility of the underlying assumption that the counterfactual time trend in the treatment group can be extrapolated from the respective trend in the control group, since a large common range (corresponding to a low level of plausibility) yields a wide range for the true value of the ATT.[2] Furthermore, for our

[1] In Public Health, the DiD design is also known as "nonequivalent control group pretest design" (Wing, Simon, and Bello-Gomez, 2018), where "nonequivalent" indicates that the group composition is not a result of randomization controlled by the researcher. In our framework, "equivalence" is used when referring to a particular type of null-hypothesis discussed in Section 3.

[2] The idea of computing and reporting the smallest common range has been suggested in Political Science by Hartman and Hidalgo (2018). However, their focus is less specific to the Difference-in-Difference framework, hence they do not provide a detailed analysis or simulations on the performance of the approach. Moreover, we

procedure based on equivalence tests, the power to reject the null hypothesis of a difference is increasing with the sample size (also see Hartman and Hidalgo, 2018). This improves upon the current practice of testing the null hypothesis of "no difference", since large samples increase the chances of rejecting the 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.

Our method can be related to several approaches in the literature on the plausibility of the PTA. For instance, Abadie, Diamond, and Hainmueller (2010) introduce a systematic way of choosing a control group as a linear combination of untreated individuals. It is then demonstrated graphically (e.g. in Abadie et al., 2010, Abadie, Diamond, and Hainmueller, 2015 or Abadie and Gardeazabal, 2003) that trends in the pre-treatment periods show a high level of similarity between the treatment and the synthetic control group. Other methods tackle the identifying assumption of parallel trends with a conditional DiD approach based on matching methods on observable covariates (e.g. Heckman, Ichimura, and Todd, 1997, Heckman, Ichimura, Smith, and Todd, 1998 or Abadie, 2005). A common feature of these approaches is that a control group is constructed such that the PTA holds exactly so that the ATT is point identified, which differs from our main approach. As we use equivalence tests, our paper is closely related to Bilinski and 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). Unlike them, we do not necessarily focus on a particular violation of the PTA. Moreover, we incorporate uncertainty about the appropriateness of the DiD design via set-identification of the average treatment effect. Other papers allow for certain deviations from exactly parallel trends. For instance, Roth and Ashesh (2020) relax the PTA by imposing restrictions based on economic knowledge which can be expressed as a set of linear inequalities on the potential differences in trends between treatment and control. They then proceed by deriving confidence sets that are valid conditional on the restriction on trend differences. While our equivalence test based procedure also allows for pre-specifying an equivalence threshold based on economic intuition, we focus on finding upper bounds for the differences in trends that are consistent with the pre-treatment data (and thus do not require specifying restrictions a priori). Chan and Kwok (2018) allow for certain violations of exactly parallel trends by augmenting the usual two-way fixed effects framework by a factor structure of the unobserved effect. However, our method is arguably more closely related to Manski and

---

suggest novel testing procedures that are specifically designed for the DiD pre-testing framework and different from the "equivalence confidence interval" of Hartman and Hidalgo (2018).

Pepper (2018), who consider a set of "bounded variation" assumptions that impose deterministic constraints on unobservable counterfactual outcomes. As a consequence, the true treatment effect is no longer point- but set-identified. Manski and Pepper then examine the sensitivity of DiD results when relaxing the PTA while imposing bounds on the "DiD Variation" (i.e. the change in group mean differences over time). A key difference to our paper is that there is no sampling uncertainty in their paper due to the use of data on the state level. Following Abadie, Athey, Imbens, and Wooldridge (2020), the remaining uncertainty in DiD estimates is thus "design based" rather "sampling based". Since there is no sampling uncertainty in Manski and Pepper (2018), selecting bounds on the DiD variation that are internally consistent with observed pre-treatment data is rather straightforward. In contrast, our paper focuses on a setting in which both design and sampling uncertainty are present and proposes a strategy that accounts for both sources of uncertainty. Finally, our paper also contributes to the literature on inference in DiD designs (e.g. Bertrand, Duflo, and Mullainathan, 2004, Donald and Lang, 2007). A key difference to previous approaches who focus on the correct measurement of the statistical insecurity in the DiD estimate (i.e. on the correct choice of standard errors) when the PTA is assumed to hold is that we focus on incorporating a measure for the uncertainty of the validity of the PTA assumption into the usual confidence intervals.

2. Pre-testing in the Difference-in-Differences framework

We focus on a repeated cross-section setup in which we observe $n_t \in \mathbb{N}$ individuals indexed by i in period $t \in \{1, \dots, T+1\}$. 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.[3] 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.[4] The observed outcome is then given by $Y_i = Y_i^0 + (Y_i^1 - Y_i^0)G_i \times D_{i,T+1}$, 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. Our object of interest is the average treatment effect on the treated

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

where X_i denotes a p-dimensional column vector of observed covariates. Since the counterfactual $\mathbb{E}[Y^0 | G_i = 1, D_{i,T+1} = 1, X_i]$ is not observed, π_{ATT} cannot be identified without further assumptions. The PTA, which ensures that in the absence of treatment both the treatment and

[3] To keep our notation simple, we pool the post-treatment periods into a single time period. We further rule out differences in treatment timing (for a discussion on differential treatment times see, for instance, Goodman-Bacon, 2018, Callaway and Sant'Anna, 2019 or Abraham and Sun, 2018).

[4] In panel data sets, the same individual may be observed in different states and periods. We demonstrate how our methodology can be applied in a panel data setting in Section 6

the control group would have experienced the same time trends between the post-treatment period $T+1$ and the "reference" period T, is given by $\Delta_{T+1}(0) - \Delta_T(0) = 0$, where

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

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 (see Callaway and Sant'Anna, 2019) given by

$$\Delta_l(0) - \Delta_l(0) = 0,\ l = 1, ..., T-1. \tag{2.1}$$

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

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

denotes the population group mean difference in period l conditional on observed characteristics. Notice that under the PTA π_{ATT} solely consists of observable quantities.

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

$$Y_i = c + G_i\alpha + \sum_{\substack{l=1 \\ l \neq T}}^{T+1} D_{i,l}\gamma_l + \sum_{\substack{l=1 \\ l \neq T}}^{T+1} \beta_l D_{i,l} \times G_i + X_i'\mu + \epsilon_i \ , \ i = 1, \ldots, n \ , \tag{2.2}$$

where c denotes a constant and μ is a p-dimensional parameter. Importantly, the group dummy G_i is time-invariant whereas the effect of the time dummies does not depend on group membership. 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. $\mathbb{E}[\epsilon_i | G_i, D_{i,1}, ..., D_{i,T-1}, D_{i,T+1}, X_i] = 0$, we should therefore expect $\beta_l = 0$ for every $l \in \{1, ..., T-1\}$. To find evidence against the plausibility of parallel trends, one could thus test for simultaneous significance of the parameters corresponding to pre-treatment periods, i.e.

$$\mathrm{H}_0 : \ \beta_1 = ... = \beta_{T-1} = 0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \exists\, l \in \{1, \ldots, T-1\} : \beta_l \neq 0. \tag{2.3}$$

As noted for instance in Roth (2020), it is however more common in applied economic research to test for individual significance, i.e. for every $l \in \{1, ..., T-1\}$ we test

$$\mathrm{H}_0 : \ \beta_l = 0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_l \neq 0. \tag{2.4}$$

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, the DiD framework is sometimes used even when H_0 in (2.4) is rejected, as some statistically significant differences are deemed negligible in a given context. Usually however, a potential threshold that quantifies what constitutes a negligible effect is insufficiently discussed. A further problematic common practice is to treat failure to reject the null hypothesis in (2.3) or (2.4) 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 (2.3) or (2.4) due to a lack of statistical power. Since the DiD framework is applied as if the PTA was true, one will then typically obtain biased estimates of the ATT. Roth (2020) further points out that since the group mean difference of the reference period is involved both in the pre-test and in the estimate of π_{ATT}, the bias is more severe conditional on not detecting an existing difference in trends. The latter aspect raises additional concerns of a "publication bias" as articles using a DiD identification argument are more likely to be deemed publishable when a test of (2.3) or (2.4) could not detect evidence against the PTA. Finally, the current approach does not make full use of the information present in the data. Commonly, the DiD framework is not applied if H_0 in (2.4) is rejected in at least one pre-treatment period since a violation of the PTA and consequentially a bias in the estimated ATT seems likely. If one however knew an "upper bound" for the extend to which the trends in treatment and control differ, one could use this information to correct the bias and thus recover useful information from a DiD analysis. As we argue in the next section, 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.

3. Testing for equivalence

Equivalence testing is well known in biostatistics (see Berger and Hsu, 1996 or Wellek, 2010). While it has recently been considered in the statistical literature for the analysis of structural breaks (e.g. Dette and Wied, 2014, Dette and Wu, 2019, Dette, Kokot, and Aue, 2020 or Dette and Wu, 2020), it is less frequently used in econometrics. To illustrate this concept in the present context, first notice that $\beta_l = \Delta_l - \Delta_T$, i.e. β_l measures the change in group mean differences between period l and the reference period conditional on a set of observed covariates. Thus, $\beta_l = 0$ signifies the absence of temporary shocks in periods l and T that only affect either treatment or control after controlling for observed regressors. 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. Thus, 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 three testing procedure that explicitly require finding evidence *in favor* of the comparability of both groups. Each of the tests yields an upper bound $\mathcal{U} \geq 0$ for changes in the group mean differences in the pre-treatment periods relative to the reference 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 $\mathcal{T}$ below which changes in the group mean differences over time are deemed negligible. On then applies our equivalence testing procedures and compares the corresponding upper bounds to the pre-specified threshold. Rejecting an equivalence test for $\mathcal{U} \leq \mathcal{T}$ at α level of significance then implies that deviations from parallel trends in the pre-treatment periods are negligible relative to the threshold $\mathcal{T}$ 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 $\mathcal{T}$ and sufficient data to support the assumption of negligible violations of the PTA pre-treatment.

However, in contrast to studies in other fields such as biology or medicine where previous experience based on clinical trials may be available to justify a threshold $\mathcal{T}$, such knowledge is rarely at hand in economic studies. It is therefore often difficult to objectively argue that a certain extend of violations of the PTA can be ignored in practice. In the remainder of the paper, we thus use the new procedures for testing equivalence developed in this paper to find the smallest upper bound $\mathcal{U}^*$ for pre-treatment changes in group mean differences. Under the assumption that the same upper bound also holds for the change in group mean differences between the reference and the post-treatment period, we can set-identify the true ATT. Formalizing this idea, we thus replace the assumption of exactly parallel trends in the absence of treatment by a "common range" assumption, i.e.

$$|\Delta_{T+1}(0) - \Delta_T(0)| \leq \mathcal{U}^*. \tag{3.1}$$

Note that this is clearly a weaker requirement than that of exactly parallel trends.[5] Since the counterfactual difference in group means is no longer given by a single number, we consequently obtain an interval of possible values of the true ATT. From (3.1), it follows that $\pi_{ATT} \in \mathrm{I}^{\mathcal{U}^*}_{ATT} = (\pi^L_{ATT}, \pi^U_{ATT})$, where $\pi^L_{ATT} := \Delta_{T+1} - \Delta_T - \mathcal{U}^*$ and $\pi^U_{ATT} := \Delta_{T+1} - \Delta_T + \mathcal{U}^*$, so that $\mathrm{I}^{\mathcal{U}^*}_{ATT}$ is

[5]As both assumptions refer to an unobservable counterfactual outcome, it is unfortunately not possible to test either assumption directly. Hence, the plausibility of exactly parallel trends or a common range has to be determined depending on the context.

an interval of length $2\mathcal{U}^*$.[6] If the PTA does not hold or its plausibility cannot be justified due to insufficient data, the uncertainty about the comparability of treatment and control in the absence of treatment is directly reflected by a wide interval of possible values of the true ATT.[7]

We develop three methods for determining a plausible upper bound $\mathcal{U}^*$. We start with a discussion of the maximum absolute change of the group mean difference in the pre-treatment periods relative to the reference period. More precisely, for a given level of significance α we find the smallest value $\delta > 0$ denoted as δ^* such that the null hypothesis in

$$\mathrm{H}_0 : \max_{l \in \{1,\dots,T-1\}} |\beta_l| > \delta \qquad \text{vs.} \qquad \mathrm{H}_1 : \max_{l \in \{1,\dots,T-1\}} |\beta_l| \leq \delta \tag{3.2}$$

is rejected. Since we are now controlling the type I error, this implies that with probability of at least $1 - \alpha$, an upper bound for the absolute change in group mean differences in the pre-treatment periods relative to the reference period is given by δ^*. If sufficient pre-treatment periods are considered, it is natural to assume that δ^* also provides an upper bound for the absolute change from the group mean difference in the post-treatment period in the absence of treatment to the group mean difference in the reference period. A potential drawback of testing (3.2) is that the common range tends to be wide and thus the procedure is rather conservative in practice.[8] In order to mitigate this problem, we consider two further alternatives of finding an upper bound. For instance, if many pre-treatment time periods are available, it may be sensible to consider the average deviation from the group mean difference in the reference period

$$\bar{\beta} := \frac{1}{T-1} \sum_{t=1}^{T-1} \beta_t. \tag{3.3}$$

Instead of testing (3.2), one may then find bounds on the average deviation from the group mean difference in the reference period by testing

$$\mathrm{H}_0 : |\bar{\beta}| > \tau \qquad \text{vs.} \qquad \mathrm{H}_1 : |\bar{\beta}| \leq \tau \tag{3.4}$$

and choosing the bound τ^* as the smallest τ for which H_0 in (3.4) is rejected. As compared to (3.2), testing the hypothesis in (3.4) can have multiple advantages in certain situations of interest. For instance, adding time periods in (3.2) can only yield wider intervals for the possible values of the true ATT. While this leads to plausible bounds on the ATT, the procedure may

[6] In some contexts, the direction of the effect may be known. In such instances, the identified set can be expressed as $\mathrm{I}^{\mathcal{U}^*}_{ATT} = (\pi^L_{ATT}, \beta_{T+1})$ or $\mathrm{I}^{\mathcal{U}^*}_{ATT} = (\beta_{T+1}, \pi^R_{ATT})$

[7] In Manski and Pepper (2018), the "DiD-variation" is bounded such that the bound is consistent with the pre-treatment data. However, their analysis uses data on the state level which removes sampling variation so that the upper bounds can be made consistent with pre-treatment data in a rather straightforward way. We similarly choose the bounds on changes in the group mean differences consistent with the pre-treatment data in a setting that includes both design and sampling uncertainty.

[8] Related criticism of equivalence tests can for instance be found in Bilinski and Hatfield (2018).

become too conservative for many pre-treatment periods. One reason is that the tests for (3.2) presented here are based on the intersection-union principle, which leads to increasingly conservative tests as the number of parameters ($= T-1$) increases. Since averaging reduces the dimension of the parameter to be tested to one irrespective of the number of pre-treatment time periods, this problem is overcome by (3.4). On the other hand, $\overline{\beta}$ and thus τ^* can be small in situations in which the summands β_t in (3.3) are large in absolute value but have different signs. In practice, one should therefore be conscious about potential cancellation effects in $\overline{\beta}$ whenever one suspects temporary shocks to the treatment or control group that switch direction between time periods.

Therefore, as a further alternative to (3.4), we consider the average squared deviation from the group mean difference in the reference period

$$\bar{\beta}_{sq} := \frac{1}{T-1}\sum_{t=1}^{T-1}\beta_t^2.$$

One then tests the hypotheses

$$\mathrm{H}_0 : \bar{\beta}_{sq} > \eta \qquad \text{vs.} \qquad \mathrm{H}_1 : \bar{\beta}_{sq} \leq \eta \tag{3.5}$$

and chooses η^* as the smallest η for which H_0 in (3.5) can be rejected. As we illustrate in our simulations, this test typically yields narrow intervals for the true ATT when the PTA holds and sufficient data is available. Thus, in cases in which the PTA is plausible due to theoretical considerations, following our procedure using (3.5) is a simple way of capturing the remaining uncertainty in the data about the comparability of treatment and control without sacrificing too much informative value about the true ATT due to overly conservative common ranges. Moreover, our test procedures can be useful in analyzing the nature of potential violations of the PTA, as we illustrate in our examples and simulations.

3.1. **Implementing equivalence tests.** We now focus on developing the test statistics for the hypotheses in (3.2), (3.4) and (3.5) which can be applied in model (2.2). To formalize the necessary assumptions, we first introduce the random vector

$$W_i := \left(1, G_i, D_{i,1}, \ldots, D_{i,T-1}, D_{i,T+1}, G_1 \times D_{i,1}, \ldots, G_i \times D_{i,T-1}, G_i \times D_{i,T+1}, X_i^\top\right)^\top$$

$(i = 1, \ldots, n)$ and the parameter

$$\theta := (c, \alpha, \gamma_1, ..., \gamma_{T-1}, \gamma_{T+1}, \beta_1, ..., \beta_{T-1}, \beta_{T+1}, \mu^\top)^\top \in \mathbb{R}^{2T+2+p}. \tag{3.6}$$

With these notations we can write model (2.2) in the form $Y_i = W_i^\top\theta + \varepsilon_i$, and the least squares estimator $\hat{\theta}$ is given by

$$\hat{\theta} = (\frac{1}{n}\sum_{i=1}^n W_i W_i^\top)^{-1}\frac{1}{n}\sum_{i=1}^n W_i Y_i = \theta + (\frac{1}{n}\sum_{i=1}^n W_i W_i^\top)^{-1}\frac{1}{n}\sum_{i=1}^n W_i\varepsilon_i. \tag{3.7}$$

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 vector $(D_{i,1}, \dots, D_{i,T+1})^\top$ 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 $\varepsilon_1 \dots, \varepsilon_n$ are independent samples of independent identically distributed random variables.
(4) The matrix $\Gamma = \mathbb{E}[W_i W_i^T]$ exists and is positive definite. $\mathbb{E}[\varepsilon_i] = \mathbb{E}[\varepsilon_i^2]$ exists and is positive.

Under these assumptions, standard arguments show that the estimate $\hat{\theta}$ in (3.6) is consistent for θ. Let further $\hat{\beta} := (\hat{\beta}_1, ..., \hat{\beta}_{T-1}, \hat{\beta}_{T+1})^\top$ denote the OLS estimator of the parameter $\beta := (\beta_1, ..., \beta_{T-1}, \beta_{T+1})^\top$ in model (2.2), then it follows that

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{N}(0, \Sigma), \tag{3.8}$$

where $N(0,\Sigma)$ denotes a T-dimensional normal distribution with mean vector $0 \in \mathbb{R}^T$ and covariance matrix $\Sigma = (\Sigma_{ij})_{i,j=1,\dots,T}$ and $n := \sum_{t=1, t\neq T}^{T+1} n_t$ denotes the total sample size. Note that $\beta = (0, \dots, 0, \pi_{ATT})^\top$ if and only if the PTA is satisfied. Based on the asymptotic normality of the OLS estimator in (3.8), we test the three different hypotheses of equivalence as follows.[9]

(1) To describe the test for the hypotheses in (3.2) we first 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 $\mathrm{N}_1(\beta_1, \Sigma_{11}/n)$, we can reject the null hypothesis

$$\mathrm{H}_0 : \ |\beta_1| > \delta \qquad \text{vs.} \qquad \mathrm{H}_1 : |\beta_1| \leq \delta$$

for small values of $|\hat{\beta}_1|$. To be precise, recall that for a normally distributed random variable $Z \sim \mathrm{N}(\mu, \sigma^2)$ the distribution of $|Z|$ is called folded normal distribution, i.e. $|Z| \sim \mathrm{N}_F(\mu, \sigma^2)$. We now propose to reject the null hypothesis in (3.2), whenever

$$|\hat{\beta}_1| \leq \mathcal{Q}_{\mathrm{N}_F(\delta, \hat{\Sigma}_{11}/n)}(\alpha), \tag{3.9}$$

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 the common estimator of the matrix Σ in (3.8). It is shown in Appendix A that this test is consistent, has asymptotic level α and is (asymptotically) uniformly most powerful for testing the hypothesis in (3.2)

[9] As discussed 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 as in Section 6.

in the case $T = 2$. For $T > 2$, we apply the idea of intersection-union tests outlined in Berger and Hsu (1996) and reject the null hypothesis in (3.2), whenever

$$|\hat{\beta}_t| \leq \mathfrak{Q}_{\mathrm{N}_F(\delta, \hat{\Sigma}_{tt}/n)}(\alpha) \ \forall t \in \{1, \ldots, T-1\}. \tag{3.10}$$

As pointed out before, it is well-known that testing procedures based on the intersection-union principle tend to be rather conservative (see Berger and Hsu, 1996, among others), which is further confirmed by Table 1 in Section 5.

(2) Next, we consider the hypothesis in (3.4) for some fixed $\tau > 0$. Writing $\hat{\beta}^{(T-1)} := (\hat{\beta}_1, \ldots, \hat{\beta}_{T-1})^\top$ so that $\hat{\beta}^{(T-1)}$ denotes the sub-vector which extracts the coordinates in the positions $T+2, \ldots, 2T$ from the vector $\hat{\theta}$, a test can be constructed by first computing the statistic

$$\bar{\hat{\beta}}^{(T-1)} := \frac{1}{T-1} \sum_{t=1}^{T-1} \hat{\beta}_t = \mathbb{1}^\top \hat{\beta}^{(T-1)}/(T-1),$$

where $\mathbb{1} = (1, \ldots, 1)^\top \in \mathbb{R}^{T-1}$. Note that it follows from (3.8) that

$$\sqrt{n}\mathbb{1}^\top(\hat{\beta}^{(T-1)} - \beta^{(T-1)}) \to \mathrm{N}(0, \mathbb{1}^\top \Sigma \mathbb{1})).$$

Consequently, based on the discussion in the previous paragraph, we propose to reject the null hypothesis in (3.4), whenever

$$|\bar{\hat{\beta}}^{(T-1)}| \leq \mathfrak{Q}_{\mathrm{N}_F(\tau, \hat{\sigma}^2)}(\alpha), \tag{3.11}$$

where $\hat{\sigma}^2 = \mathbb{1}^\top \hat{\Sigma} \mathbb{1}/(n(T-1)^2)$. While this test maintains its nominal level for every $T \geq 2$, a potential downside (as mentioned before) of the hypothesis (3.4) is that vectors with large absolute entries and opposite signs are classified through the mean $\bar{\beta}$ as small. In practice, it can therefore be informative to combine this test with a test for the hypotheses in (3.5) which does not encounter this issue.

(3) In order to construct a pivot test for the hypotheses (3.5), recall the definition of the OLS estimator $\hat{\theta}$ in (3.7) and let, with a slight abuse of notation, $\varepsilon > 0$ denote a small positive constant. For $\lambda \in [\varepsilon, 1]$, define

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

as the OLS estimator for the parameter θ in (3.6) from the sample $(W_1, Y_1), \ldots, (W_{\lfloor n\lambda \rfloor}, Y_{\lfloor n\lambda \rfloor})$, such that for sufficiently large sample sizes $\hat{\theta}(\lambda)$ is well defined. Next, define

$$\hat{\beta}^{(T-1)}(\lambda) = (\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_{T-1}(\lambda))^\top$$

as a sub-vector of $\hat{\theta}(\lambda)$ extracting the coordinates in the positions $T+2, \ldots, 2T+2$. Further notice that $\hat{\beta}^{(T-1)}(1)$ is the OLS estimator of $\beta^{(T-1)}$ based on the full sample,

i.e. $\hat{\beta}^{(T-1)}(1) = \hat{\beta}^{(T-1)}$. We now define

$$\hat{M}_n := \frac{\frac{1}{T-1}\|\hat{\beta}^{(T-1)}(1)\|^2 - \frac{1}{T-1}\|\beta^{(T-1)}\|^2}{\hat{V}_n}, \tag{3.12}$$

where $\|\cdot\|$ denotes the euclidean norm on $\mathbb{R}^{T-1}$,

$$\hat{V}_n = \frac{1}{T-1}\Big(\int_\varepsilon^1 (\|\hat{\beta}^{(T-1)}(\lambda)\|^2 - \|\hat{\beta}^{(T-1)}(1)\|^2)^2 \nu(d\,\lambda)\Big)^{1/2} \tag{3.13}$$

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

Theorem 3.1. *If Assumption 3.1 is satisfied and $\beta^{(T-1)} \neq 0$, then the statistic $\hat{M}_n$ defined in* (3.12) *converges weakly with a non-degenerate limit distribution, that is*

$$\hat{M}_n \xrightarrow{d} \mathbb{W} := \frac{\mathbb{B}(1)}{\big(\int_\varepsilon^1 (\mathbb{B}(\lambda)/\lambda - \mathbb{B}(1))^2 \nu(d\lambda)\big)^{1/2}}, \tag{3.14}$$

where $\{\mathbb{B}(\lambda)\}_{\lambda\in[\varepsilon,1]}$ is a Brownian motion on the interval $[\varepsilon, 1]$.

It follows from the proof of Theorem 3.1 that the statistic $\frac{1}{T-1}\|\hat{\beta}^{(T-1)}\|^2$ is a consistent estimator of $\overline{\beta}_{sq} = \frac{1}{T-1}\|\beta^{(T-1)}\|^2$. Therefore, we propose to reject the null hypothesis H_0 in (3.5), whenever

$$\frac{1}{T-1}\big\|\hat{\beta}^{(T-1)}(1)\big\|^2 = \frac{1}{T-1}\sum_{t=1}^{T-1}\hat{\beta}_t^2 \leq \eta + \mathcal{Q}_{\mathbb{W}}(\alpha)\hat{V}_n, \tag{3.15}$$

where $\mathcal{Q}_{\mathbb{W}}(\alpha)$ is the α-quantile of the distribution of the random variable $\mathbb{W}$ on the right-hand side of (3.14). Note that these quantiles can be easily obtained by simulation. The following result shows that this decision rule defines a valid test for the hypotheses in (3.5).

Theorem 3.2. *If Assumption 3.1 is satisfied, then the test defined by* (3.15) *is a consistent asymptotic level α-test for the hypothesis in* (3.5)*, that is*

$$\lim_{n\to\infty} \mathbb{P}_{\bar{\beta}_{sq}}\Big(\frac{1}{T-1}\|\hat{\beta}^{(T-1)}(1)\|^2 \leq \eta + Q_{\mathbb{W}}(\alpha)\hat{V}_n\Big) = \begin{cases} 0, & \text{if} \quad \bar{\beta}_{sq} > \eta \\ \alpha, & \text{if} \quad \bar{\beta}_{sq} = \eta \\ 1, & \text{if} \quad \bar{\beta}_{sq} < \eta \end{cases}$$

Remark 3.1. Notice that in practice one chooses ν as a discrete distribution which makes the evaluation of the integrals in (3.13) 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 $\hat{V}_n^2$ in (3.13) simplifies to

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

This measure is also used in the simulation study in Section 5, where we analyze the finite sample properties of the different procedures.

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

(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.2 shows that similar results can be obtained in the case of dependent data. More precisely, for a symmetric $d \times d$ matrix A let $\mathrm{vech}(A)$ denote the $d(d+1)/2$-dimensional vector that stacks the columns of the matrix A below the diagonal in a vector, where $d := 2T + 2 + p$. Let K denote a non-singular $d \times d$-matrix and let $\mathbb{B}$ be a d-dimensional vector of independent Brownian motions. The d-dimensional time series $\{(W_i, \varepsilon_1)\}_{i=1}^n$ is stationary and the sequential process

$$\left\{ \sqrt{n} \begin{pmatrix} \frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} W_i \varepsilon_i \\ \mathrm{vech}(\frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} W_i W_i^T - \Gamma) \end{pmatrix} \right\}_{\lambda \in [\varepsilon, 1]} \rightsquigarrow \left\{ K \frac{\mathbb{B}(\lambda)}{\lambda} \right\}_{\lambda \in [\varepsilon, 1]} \tag{3.16}$$

converges weakly 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 (3.16) have been proved for many dependence concepts in the literature (such as different types of mixing or physical dependence; see, for instance, Merlevède, Peligrad, and Utev, 2006 and the references in therein).

(2) Similarly, note that 3.1(b), 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 $D_i = (D_{i,1}, \ldots, D_{i,T+1})^\top$ is not random and given by $(1, \ldots, 1)^\top$. Moreover, panel data with missing observation can be also modeled using a random vector $D_i = (U_{i,1}, \ldots, U_{i,T-1}, U_{i,T+1})^\top$ where $U_{i,1}, \ldots, U_{i,T-1}, U_{i,T+1}$ are independent Bernoulli variables with success probabilities $p_1, \ldots, p_{T-1}, p_{T+1}$, respectively (here $1 - p_t$ represents the probability that an observation for the i-th object is not available for time t).

4. Equivalence testing in practice

We begin this section with reviewing the credibility of the PTA in situations of applied interest. We then proceed by discussing the effect of violations of the PTA on our equivalence tests. Moreover, we present a simple approach of obtaining confidence intervals that take into account the two sources of uncertainty about the treatment effect estimate, namely the sampling variation of the usual DiD estimator and the uncertainty about the DiD design as epressed by the common range defined by our equivalence tests.

4.1. **Failure of the PTA.** As shown in Lechner (2011), the PTA is only credible if variables that lead to unconditional differences between the composition of treatment and control are conditioned on. As Lechner further points out, the presence of omitted variables however does not necessarily lead to biased estimates of the ATT. The latter nonetheless requires some strong restrictions on the type of omitted variable. For instance, a variable U can be ignored if the PTA would hold conditional on U and the distribution of U does not depend upon group membership. Moreover, the unobserved variable may safely be ignored when its effect on the potential outcomes varies across groups but is constant over time.

In practice, both cases are often too restrictive. As noted for instance in Heckman and Smith (1999), DiD estimation of treatment effects may be problematic in the presence of self-selection into treatment that is not accounted for in the estimation procedure. Moreover, estimated treatment effects may be biased if individuals adapt their behavior in anticipation of future treatment (Lechner, 2011). Ashenfelter (1978) further observed that participants in public training programs often suffer from a larger pre-treatment drop in earnings as compared to non-participants, a phenomenon now known as "Ashenfelter's dip". As pointed out in Heckman and Smith (1999), a pre-program dip may lead to overstated or understated treatment effects, depending on the nature of the decline in pre-treatment earnings.

4.2. **Testing for equivalence under violations of the PTA.** In order to formalize the violations of the PTA discussed in the previous paragraph, we now consider the model in (2.2) 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

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

where $\tilde{\varepsilon}_i = Z_i^\top \nu + \varepsilon_i$ and ε_i satisfies Assumption 3.1(1). The OLS estimator now becomes

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

As $\mathbb{E}[W_i\varepsilon_i] = 0$ under Assumption 3.1, it is easy to see that the OLS estimator is only consistent as $n \to \infty$ if $\mathbb{E}[W_i Z_i^\top] = 0$. Notice that in the presence of Z_i we have $\beta_l = \Delta_l - \Delta_T - (\Delta_l^Z - \Delta_T^Z)$, where

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

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

$$\rho = \Gamma^{-1}\mathbb{E}[W_i Z_i^\top \nu]$$

is the omitted variable bias and the matrix Γ is defined in Assumption 3.1. Therefore, when the true effect of the treatment prior to the treatment is zero, i.e. $\beta_l = 0$ for $l = 1, ..., T-1$, our equivalence tests implicitly yield an upper bound for the omitted variable bias ρ, which in certain situations (see for instance Example 4.1 below) can be used to correct the estimate of the true treatment effect $\hat{\pi}_{ATT}$.

4.3. **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 state 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 reference period then becomes

$$\Delta_{T+1} - \Delta_T = \beta_{T+1} + \Delta^Z_{T+1} - \Delta^Z_T = \beta_{T+1} - \nu,$$

since $\Delta^Z_{T+1} = 0$ and $\Delta^Z_T = \nu$. 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. A similar result holds if more than one pre-treatment period is available, since $\text{plim}_{n\to\infty} \hat{\beta}_l = -\nu_1$ for all $l = 1, ..., T-1$. Thus, the maximum change in the group mean differences between the pre-treatment periods and the reference period equals the average change. Consequently, testing the null hypotheses in (3.2) and (3.4) at level of significance α using the statistics in (3.10) and (3.11) respectively yields an upper bound on the absolute value of the omitted variable bias that holds with probability $1 - \alpha$.

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 variables

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 $\Delta^Z_{T+1} - \Delta^Z_T = (T+1)\psi - T\psi = \psi$. Therefore, the change in the (conditional) mean difference between the groups from the post-treatment period relative to the reference period differs from the true ATT by ψ. Moreover, $\delta_l - \Delta_T = \beta_l + \psi(l - T) = \psi(l - T)$, so that $|\hat{\beta}_l|$ will typically increase with $|l - T|$. Thus, δ^*, τ^* and η^* 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 η^* 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.

4.4. **Accounting for design uncertainty.** Our approach is targeted at applications in which the estimate of the true ATT contains two sources of uncertainty. The first source is the usual statistical variation of the OLS estimator, which can be taken into account by using appropriate estimators of the standard errors (Bertrand et al., 2004).[10] The second source is the uncertainty about the appropriateness of the DiD design, which is usually assessed by testing (2.4) but assumed away if H_0 cannot be rejected. Instead of ignoring the second source of uncertainty, we propose a simple method to construct confidence intervals that take into account both the sampling and the design uncertainty. Let $\mathrm{CI}^{\hat{\beta}_{T+1}} = (\mathrm{CI}_L^{\hat{\beta}_{T+1}}, \mathrm{CI}_R^{\hat{\beta}_{T+1}})$ where $\mathrm{CI}_L^{\hat{\beta}_{T+1}}$ and $\mathrm{CI}_R^{\hat{\beta}_{T+1}}$ denote the left and right endpoint of the confidence interval obtained by inverting the usual t-statistic with the appropriately chosen standard error. The interval $\mathrm{I}^{\mathcal{U}^*}_{ATT}$ with $\mathcal{U}^* \in \{\delta^*, \tau^*, \eta^*\}$ provides a measure of "confidence" in the PTA as indicated by our equivalence tests. A simple confidence interval that includes both the design and the sampling uncertainty can be obtained by combining the two previous intervals by considering the interval $\tilde{\mathrm{CI}}^{\mathcal{U}^*} = (\tilde{\mathrm{CI}}_L^{\mathcal{U}^*}, \tilde{\mathrm{CI}}_R^{\mathcal{U}^*})$ with $\tilde{\mathrm{CI}}_L^{\mathcal{U}^*} := \pi^L_{ATT} - \mathrm{CI}_L^{\hat{\beta}_{T+1}}$ and $\tilde{\mathrm{CI}}_R^{\mathcal{U}^*} := \pi^R_{ATT} + \mathrm{CI}_R^{\hat{\beta}_{T+1}}$. Naturally, the "combined confidence interval" $\tilde{\mathrm{CI}}^{\phi^*}$ is a more conservative confidence interval as compared to $\mathrm{CI}^{\hat{\beta}_{T+1}}$ as taking into account the design uncertainty can only lead to wider confidence intervals. It is however demonstrated in Section 5 below that when the PTA is violated, $\mathrm{CI}^{\hat{\beta}_{T+1}}$ can severely undercover π_{ATT} while the coverage probability of π_{ATT} of $\tilde{\mathrm{CI}}^{\mathcal{U}^*}$ is closer to the nominal level and may thus yield a more reliable confidence interval for π_{ATT}. In this simple way, we obtain a confidence interval that is "honest" in the sense described in Roth and Ashesh (2020) as long as the common range assumption (3.1) holds and the appropriate standard errors are chosen (Bertrand et al., 2004). When conducting significance tests, p-values that account

[10] Our approach is not appropriate for applications based on "aggregate data" (Abadie et al., 2010) on the population level such as for instance in Manski and Pepper (2018), as our tests are designed to yield likely upper bounds for differential trends on the aggregate level from non-aggregate data.

for both sources of uncertainty can be computed in the same spirit by considering the usual t-statistic with $\hat{\beta}_{T+1}$ in the numerator being replaced by the appropriate endpoint of $\mathrm{I}_{ATT}^{\mathcal{U}^*}$.

5. Simulations

In order to investigate the small sample properties of our procedure, 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. Thus, the total sample size n is given by $\sum_{t=1}^{T+1} n_t$. In all our simulations we set $p_G = \frac{1}{2}$ and $p_{D_l} = \frac{1}{T+1}$ so that 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 ϵ_i independently from a standard normal distribution. The data for X_i is independently drawn from a normal distribution with mean 1 and standard deviation 2. The number of Monte Carlo iterations is 50000 in all simulation results reported.

In an initial step, we investigate the level of each of the three tests we propose. To do so, we choose the level of significance $\alpha = 5\%$ and let $\beta_t = 1$ for $t = 1, ..., T-1$. We then set $\delta = \tau = \eta = 1$, where δ, τ and η are the respective equivalence bounds in (3.10), (3.11) and (3.15). The results are presented in Table 1.

In the following scenarios, we report δ^*, τ^* and η^* as the smallest values such that the corresponding null hypothesis in (3.2), (3.4) and (3.5) can be rejected at level $\alpha = 5\%$. Further, we consider empirical frequencies for the true treatment effect to lie in the intervals discussed in Section 4.4. We report the usual 95% confidence interval $\mathrm{CI}^{\hat{\beta}_{T+1}}$, which captures the sampling uncertainty in $\hat{\beta}_{T+1}$, the "common range" $\mathrm{CI}^{\mathcal{U}^*} = (-\mathcal{U}^*, \mathcal{U}^*)$ with $\mathcal{U}^* \in \{\delta^*, \tau^*, \eta^*\}$ obtained by testing (3.2), (3.4) and (3.5) at 5% level of significance capturing the design uncertainty and the combined confidence interval $\tilde{\mathrm{CI}}^{\mathcal{U}^*}$ capturing both sampling and design uncertainty. Finally, we report the width of the common ranges defined by each of our three test procedures relative to the width of the confidence interval of the estimated ATT as $=\mathrm{R}(\mathrm{CI}^{\mathcal{U}^*}) := 2\mathcal{U}^*/(\mathrm{CI}_R^{\hat{\beta}_{T+1}} - \mathrm{CI}_L^{\hat{\beta}_{T+1}})$. This helps assessing the relative increase in uncertainty about the ATT when taking into account the uncertainty about the appropriateness of the DiD methodology. In all our scenarios we set $\beta_{T+1} = 0$ so that the treatment has no effect. We then investigate how often a non-existing effect is detected with our methodology as compared to the usual confidence interval that ignores design uncertainty.

Tables 3 and 4 show the effect of our approach on the estimation of the treatment effect under the PTA. While Table 3 presents the results for all simulated cases, Table 4 focuses on only those cases in which $\hat{\beta}_l$ is statistically insignificant at 5% level for every $l = 1, ..., T-1$. 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 $\in \{\frac{1}{4}, \frac{1}{2}\}$ and variance 1. The results are given in Table 5 and 6. 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 7 and 8.

5.1. **Simulation results.** As shown in Table 1, the test in (3.2) maintains its nominal level for $T = 2$ but becomes conservative for larger values of T. This phenomenon is well-known for tests constructed with the union-intersection principle (Berger and Hsu, 1996). The test in (3.4) approximately keeps the desired level for every T even for small samples, whereas the test in (3.5) is slightly over-rejecting when $n_t = 100$ but keeps its nominal level in larger samples.

	$n_t = 100$				$n_t = 1000$			
Test	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
(3.2)	0.0503	0.0051	0.0007	0.0003	0.0477	0.0050	0.0008	0.0002
(3.4)	0.0503	0.0496	0.0503	0.0504	0.0477	0.0484	0.0483	0.0501
(3.5)	0.0994	0.0797	0.0773	0.0739	0.0607	0.0571	0.0570	0.0572

TABLE 1. Rejection frequencies of the tests in (3.2), (3.4) and (3.5) for $\beta_t = 1$, $t = 1, ..., T-1$ and $\delta = \gamma = \eta = 1$ at nominal level of significance $\alpha = 5\%$.

The fact that the test in (3.2) becomes very conservative may explain why δ^* is increasing in T for all sample sizes and in all simulation setups. As can be seen from Table 3, the latter is true even when the PTA holds. One of the reasons for this behavior is that due to its construction, the value of δ^* is largely determined by the maximal variation in the components of $(\hat{\beta}_1, ..., \hat{\beta}_{T-1})$. Naturally, CI^{δ^*} and particularly $\tilde{\text{CI}}^{\delta^*}$ contains the true ATT in almost 100% of all cases, especially as T increases. However, as is indicated by $\text{R}(\text{CI}^{\delta^*})$, this comes at the cost of a much larger total confidence interval for the ATT, as under the PTA $\text{CI}^{\hat{\beta}_{T+1}}$ keeps the nominal level of 95%. For instance, taking into account the uncertainty about the appropriateness of the DiD framework using (3.2) more that doubles the width of the usual confidence interval for the ATT for every value of T. In comparison, the tests in (3.4) and (3.5) perform much better when the sample is sufficiently large. As can be seen in Tables 3 and 4, the values for τ^* and η^* tend to decrease with T. In particular the test procedure in (3.5) performs very well, as the width of the usual confidence interval is increased by less that a third when taking into account the uncertainty about the DiD framework. Nevertheless, the total confidence interval based on η^* still covers the true ATT in more than 98% of all cases. 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 4, as an increase in available pre-treatment periods increases the chance of incorrectly rejecting the DiD framework under the PTA. Thus, instead of rejecting a DiD analysis in an application the PTA is well-founded due to a significant pre-treatment parameter estimate, one could adapt the total confidence interval approach based on η^* without much loss of precision. A similar observation can be made in the presence of a linear time trend as shown in Tables aaa. 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 temporary shock as in Ashenfelter's dip, $\hat{\beta}_{T+1}$ is biased and the usual confidence interval $\text{CI}^{\hat{\beta}_{T+1}}$ may severely undercover the true ATT, as can be seen from Table 6. Notice that the bias corresponds to the negative of the mean of the temporary shock to the treatment group, as is expected by the analysis in Example 4.1. Again, δ^* increases in T, whereas τ^* and η^* remain stable or slightly decrease in T, as variation in the pre-treatment components of $\hat{\beta}$ is "smoothed out" with more pre-treatment periods available. Unlike under the PTA, $\text{R}(\text{CI}^{\tau^*})$ and $\text{R}(\text{CI}^{\eta^*})$ are rather large due to the (correctly detected) uncertainty about the DiD framework that is larger than the uncertainty about the precision of $\hat{\beta}_{T+1}$. Taking this uncertainty into account, $\tilde{\text{CI}}^{\tau^*}$ covers the ATT in all simulated cases, whereas $\tilde{\text{CI}}^{\eta^*}$ covers π_{ATT} in more than 85% of all simulated cases, which vastly improves upon the coverage of $\text{CI}^{\hat{\beta}_{T+1}}$ of less than 0.1%.

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 $\text{CI}^{\hat{\beta}_{T+1}}$ is small), $\text{CI}^{\hat{\beta}_{T+1}}$ contains the true ATT in less than 95% of the cases. If the slope of the time trend is small, the coverage of $\text{CI}^{\hat{\beta}_{T+1}}$ is however close to its nominal level, as is shown in Table 7. As expected, the coverage gets worse with a more pronounced slope of the time trends (Table 8).[11] Our methodology can be useful in two ways: First, it can help identifying the presence of a linear time trend, as τ^* and η^* 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). Second, once a time trend has been identified, $\tilde{\text{CI}}^{\tau^*}$ and $\tilde{\text{CI}}^{\eta^*}$ may still provide useful (although rather conservative) total confidence intervals of the ATT, as the empirical coverage probabilities exceed the nominal coverage level, whereas $\text{CI}^{\hat{\beta}_{T+1}}$ can severely undercover π_{ATT}.

[11] As shown in Example 4.2, the bias of $\hat{\beta}_{T+1}$ also increases in the time distance between the post-treatment period to the reference period.

6. Empirical illustration

In this section, we illustrate our approach by re-considering Difference-in-Differences analysis in Di Tella and 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. 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 difficulty of obtaining credible causal estimates from a DiD design is highlighted by Donohue, Ho, and Leahy (2013), who question the credibility of the analysis of Di Tella and Schargrodsky (2004) as the terrorist attack may also have affected the control group (i.e. blocks without Jewish institutions) therefore contaminating the DiD estimates. Instead of focusing on the validity of DiD in the post-treatment periods, we re-examine the data for differences between treatment and control before the treatment in order to understand what causal estimates could in principle be obtained by comparing the treatment and control groups in the data at hand. The main specification in Di Tella and 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 $D_{it} = 1$ if block i is treated and t refers to a post-treatment period and $D_{it} = 0$ otherwise. 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 (2.2) 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.[12] Secondly, we replace the block-specific dummies by a single group dummy and compute heteroscedasticity-robust standard errors.[13] Finally, we run separate regressions for each of the potential reference periods (i.e. for each of the pre-treatment periods) and compute the corresponding smallest upper bounds for the tests in (3.2), (3.4) and (3.5).[14] The results are summarized in Table 2 below.[15]

[12]Cluster-robust standard errors are computed using the **R** function `cluster.vcov` on the block level.

[13]White standard errors are computed using the **R** function `vcovHC` with the option "HC1".

[14]Notice that the test in (3.5) does not require an estimator of the asymptotic variance. Thus, it is not affected by the choice of standard errors.

[15]According to Assumption 3.1, if the sample size is sufficiently large, each subsample contains individuals from all time periods and groups. In practice, many data sets are "ordered" (e.g. by time) such that the first $\lfloor \lambda n \rfloor$ observations may not contain individuals observed in later time periods. Since this is the case for the data provided by Di Tella and Schargrodsky (2004), it is thus not possible to apply the test in (3.5) directly. In order to circumvent this issue, we implement our test by drawing random samples of size $\lfloor \lambda n \rfloor$ instead of using the $\lfloor \lambda n \rfloor$ first observations. To mitigate dependence on a particular draw, we repeat the procedure $B = 500$ times and report the average smallest upper bound η^*.

Estimates \ Reference period	April	May	June	July
$\hat{\pi}_{ATT}$	-0.081	-0.058	-0.121	-0.049
p-value	0.067	0.192	0.006	0.269
(3.2) (clustered)	0.147	0.128	0.158	0.104
(3.4) (clustered)	0.042	0.091	0.144	0.076
(3.2) (White)	0.127	0.158	0.158	0.135
(3.4) (White)	0.07	0.093	0.136	0.056
(3.5)	0.037	0.024	0.034	0.02

TABLE 2. Treatment effect estimates with corresponding p-values and upper bounds based on the test procedures in (3.2), (3.4) and (3.5) for different choices of the reference period.

While we should not expect large differences in the treatment effect estimates under perfectly parallel trends, we find that the choice of the reference period has a substantial effect on $\hat{\pi}_{ATT}$. Indeed, the pooled effect reported by Di Tella and Schargrodsky (2004) seems to be largely driven by a large change in differences between treated and untreated blocks between June and the post-treatment periods, whereas the corresponding changes between the remaining pre-treatment periods and the post-treatment periods are at most marginally significant.[16] We further find rather small differences between the standard errors of $\hat{\pi_{ATT}}$ obtained from the two specifications (0.044 (White) and 0.042 (clustered) independent of the reference period). Thus, Table 2 illustrates that the combined confidence intervals $\tilde{\text{CI}}^{\delta^*}$, $\tilde{\text{CI}}^{\tau^*}$ and $\tilde{\text{CI}}^{\eta^*}$ almost always contain zero, so that "no effect" is a plausible explanation for the observed data after accounting for design and sampling uncertainty.[17] These findings are in line with Donohue et al. (2013), who find that pre-treatment crime levels differ substantially between the neighborhood containing the majority of treated blocks and the remaining two neighborhoods. It is thus not surprising that our equivalence tests find little support for the comparability of treatment and control.

7. Conclusion

We have shown a way of accounting for potentially non-parallel trends by replacing the usual assumptions of exactly parallel paths by a common range based on three distinct equivalence

[16]The estimates obtained in Di Tella and Schargrodsky (2004, Table 3) by pooling the pre-treatment periods are significant and range between -0.058 and -0.081.

[17]In fact the only total confidence interval for which "no effect" is not plausible is taking June as reference period and accounting for design uncertainty through the "least squared" test in (3.5).

tests. Our tests capture the maximum, average and squared average change in group mean differences relative to the reference period and thus measure similarity between treatment and control groups. We further provide a simple way of accounting for both design and sampling uncertainty which, as compared to the standard confidence interval, is shown to yield more reliable inference when the parallel trends assumption does not hold. To illustrate our approach, we finally apply our methodology to the data provided by Di Tella and Schargrodsky (2004) and conclude that when design uncertainty is taken into account using our equivalence tests, there is no significant treatment effect.

Acknowledgements This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1) of the German Research Foundation (DFG).

References

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

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020): "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 88, 265–296.

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program," *Journal of the American statistical Association*, 105, 493–505.

——— (2015): "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59, 495–510.

Abadie, A. and J. Gardeazabal (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132.

Abraham, S. and L. Sun (2018): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Available at SSRN 3158747*.

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

Ashenfelter, O. (1978): "Estimating the effect of training programs on earnings," *The Review of Economics and Statistics*, 47–57.

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

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics*, 119, 249–275.

Bilinski, A. and L. A. Hatfield (2018): "Seeking evidence of absence: reconsidering tests of model assumptions," *arXiv preprint arXiv:1805.03273*.

——— (2020): "Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions," *Archiv*.

Callaway, B. and P. H. Sant'Anna (2019): "Difference-in-differences with multiple time periods," *Available at SSRN 3148250*.

Chan, M. K. and S. Kwok (2018): "Difference-in-Differences when Trends are Uncommon and Stochastic," *Available at SSRN 3125890*.

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

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

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

——— (2020): "Testing relevant hypotheses in functional time series via self-normalization," *Journal of the Royal Statistical Society: Series B, to appear arXiv:1809.06092*.

Di Tella, R. and E. Schargrodsky (2004): "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces After a Terrorist Attack," *American Economic Review*, 94, 115–133.

Donald, S. G. and K. Lang (2007): "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 89, 221–233.

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

Goodman-Bacon, A. (2018): "Difference-in-differences with variation in treatment timing," Tech. rep., National Bureau of Economic Research.

Hartman, E. and F. D. Hidalgo (2018): "An Equivalence Approach to Balance and Placebo Tests," *American Journal of Political Science*, 62, 1000–1013.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing selection bias using experimental data," *Econometrica*, 66, 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64, 605–654.

Heckman, J. J. and J. A. Smith (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, 313–348.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017): "The Power of Bias in Economics Research," *The Economic Journal*, 127, F236–F265.

Kahn-Lang, A. and K. Lang (2019): "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business & Economic Statistics*, 0, 1–14.

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

Manski, C. F. and J. V. Pepper (2018): "How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions," *Review of Economics and Statistics*, 100, 232–244.

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

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

Roth, J. (2020): "Pre-test with Caution Event-study Estimates After Testing for Parallel Trends," Tech. rep.

Roth, J. and R. Ashesh (2020): "An Honest Approach to Parallel Trends," Tech. rep.

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

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

Wing, C., K. Simon, and R. A. Bello-Gomez (2018): "Designing difference in difference studies: best practices for public health policy research," *Annual review of public health*, 39.

Faculty of Mathematics, Ruhr University Bochum, D-44801 Bochum, Germany.

E-mail address: holger.dette@ruhr-uni-bochum.de

School of Business and Economics, Maastricht University, 6211 LK Maastricht, The Netherlands.

E-mail address: m.schumann@maastrichtuniversity.nl

Appendix A. Mathematical proofs

A.1. **Properties of the test** (3.9)**.** For sufficiently large sample sizes the quantile $f_\alpha := Q_{\mathrm{N}_F}(\delta, \hat{\Sigma}_{11}/n)$ satisfies

$$\alpha = \mathbb{P}\big(\mid \mathrm{N}_F(\delta, \hat{\Sigma}_{11}/n) \mid \leq Q_{\mathrm{N}_F}(\delta, \hat{\Sigma}_{11}/n)\big) \approx \Phi\Big(\frac{f_\alpha - \delta}{\Sigma_{11}}\Big) - \Phi\Big(\frac{-f_\alpha - \delta}{\Sigma_{11}}\Big) \tag{A.1}$$

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

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

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 et al. (2005)).

A.2. **Proof of Theorem 3.1.** Recall that $\hat{\theta}(\lambda)$ is the OLS for the parameter θ in model (4.1) from the observations $(W_1, Y_1), \ldots, (W_{\lfloor n\lambda \rfloor}, Y_{\lfloor n\lambda \rfloor})$, that is

$$\hat{\theta}(\lambda) \;=\; \Gamma_{\lfloor n\lambda \rfloor}^{-1} \frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} W_i Y_i = \theta + \Gamma_{\lfloor n\lambda \rfloor}^{-1} \frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} W_i \varepsilon_i,$$

where the matrix Γ_k is defined by

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

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

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

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 and Wellner (1996) that

$$\big\{\sqrt{n}(\hat{\theta}(\lambda) - \theta)\big\}_{\lambda \in [\varepsilon, 1]} \rightsquigarrow \Big\{\frac{\tau \Gamma^{-1/2}}{\lambda} \mathbb{B}(\lambda)\Big\}_{\lambda \in [\varepsilon, 1]} \tag{A.3}$$

where $\mathbb{B}$ is a $2T + 2 + p$-dimensional vector of independent Brownian motions $\tau^2 = \mathrm{Var}(\varepsilon_i)$ and the symbol $\rightsquigarrow$ means weak convergence in space $(\ell^\infty[\varepsilon, 1])^{2T+2+p}$ of all $(2T + 2 + p)$-dimensional

bounded functions on the interval $[\varepsilon, 1]$. As the projections of θ on its coordinates are continuous mappings the weak convergence (A.3) and the continuous mapping theorem imply

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

where D is a $(T-1)\times(2T+2+p)$ matrix of full rank. Therefore it follows that

$$\begin{aligned} H_n(\lambda) &= \sqrt{n}\big(\|\hat\beta^{(T-1)}(\lambda)\|^2 - \|\beta^{(T-1)}\|^2\big) \\ &= \sqrt{n}\{\|\hat\beta^{(T-1)}(\lambda) - \beta^{(T-1)}\|^2 + 2(\hat\beta^{(T-1)}(\lambda) - \beta^{(T-1)})^\top \beta^{(T-1)} \\ &= 2\sqrt{n}(\hat\beta^{(T-1)}(\lambda) - \beta^{(T-1)})^\top \beta^{(T-1)} + o_{\mathbb{P}}(1) \end{aligned}$$

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

$$\big\{H_n(\lambda)\big\}_{\lambda\in[\varepsilon,1]} \rightsquigarrow \Big\{2(\beta^{(T-1)})^\top D\frac{\mathbb{B}(\lambda)}{\lambda}\Big\}_{\lambda\in[\varepsilon,1]}$$

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

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

where $\mathbb{B}_1$ is a one-dimensional Brownian motion and $\Delta > 0$ an appropriate constant. Recalling the definition of the statistic $\hat M_n$ in (3.12) and a further application of the continuous mapping theorem shows that

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

which proves the assertion.

A.3. **Proof of Theorem 3.2.** Observing the definition of $\hat M_T$ in (3.12) we obtain

$$\mathbb{P}_{\bar\beta_{sq}}\Big(\frac{1}{T-1}\|\hat\beta(1)^{(T-1)}\|^2 \le \eta + Q_{\mathbb{W}}(\alpha)\hat V_n\Big) = \mathbb{P}\Big(\hat M_T \le \frac{\eta - \bar\beta_{sq}}{\hat V_n} + Q_{\mathbb{W}}(\alpha)\Big).$$

It follows from the proof of Theorem 3.1 that $\hat V_n = O_{\mathbb{P}}(1/\sqrt{n})$. Consequently, if $\bar\beta_{sq} > 0$, assertion (3.1) follows by a simple calculation considering the three cases separately. On the other hand, if $\bar\beta_{sq} = 0$, the proof of Theorem 3.1 also shows that $\|\hat\beta^{(T-1)}(1)\|^2 = O_{\mathbb{P}}(\frac{1}{n})$ and the assertion is obvious.

Appendix B. Simulation results

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\hat{\beta}_{T+1}$	0.0000	−0.0001	0.0005	0.0008	0.0002	0.0005	-0.0002	0.0001
$\mathrm{CI}^{\hat{\beta}_{T+1}}$	0.9487	0.9512	0.9475	0.9504	0.9488	0.9496	0.9512	0.9511
δ^*	0.6428	0.8245	0.9242	0.9681	0.2018	0.2587	0.2881	0.3027
τ^*	0.6428	0.5242	0.4878	0.4747	0.2018	0.1650	0.1524	0.1489
η^*	0.6404	0.5713	0.5472	0.5366	0.0585	0.0520	0.0494	0.0488
CI^{δ^*}	0.9097	0.9959	0.9995	0.9997	0.9110	0.9958	0.9995	0.9998
CI^{τ^*}	0.9097	0.8889	0.8780	0.8746	0.9110	0.8900	0.8792	0.8742
CI^{η^*}	0.7488	0.7960	0.8156	0.8270	0.3934	0.3844	0.3807	0.3767
$\tilde{\mathrm{CI}}^{\delta^*}$	0.9985	1.0000	1.000	1.0000	0.9987	1.0000	1.0000	1.0000
$\tilde{\mathrm{CI}}^{\tau^*}$	0.9985	0.9993	0.9994	0.9995	0.9987	0.9992	0.9996	0.9995
$\tilde{\mathrm{CI}}^{\eta^*}$	0.9969	0.9987	0.9991	0.9994	0.9864	0.9874	0.988	0.9887
$\mathrm{R}(\mathrm{CI}^{\delta^*})$	1.1456	1.4712	1.6503	1.7294	1.1497	1.4744	1.6417	1.7246
$\mathrm{R}(\mathrm{CI}^{\tau^*})$	1.1456	0.9350	0.8704	0.8473	1.1497	0.9400	0.8683	0.8484
$\mathrm{R}(\mathrm{CI}^{\eta^*})$	1.1405	1.0179	0.9758	0.9572	0.3334	0.2966	0.2817	0.2781

Table 3. Estimation and test performance under the PTA at nominal level of significance $\alpha = 5\%$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.9504	0.8755	0.7723	0.6967	0.9493	0.8737	0.7688	0.6971
$\hat{\beta}_{T+1}$	0.0001	-0.0008	-0.0005	0.0013	0.0000	-0.0003	-0.0003	0.0001
$\text{CI}^{\hat{\beta}_{T+1}}$	0.9561	0.9642	0.9737	0.9785	0.9568	0.9648	0.9739	0.9763
δ^*	0.6182	0.7807	0.8545	0.8862	0.1937	0.2443	0.2672	0.2769
τ^*	0.6182	0.4864	0.4304	0.4055	0.1937	0.1523	0.1350	0.1276
η^*	0.6028	0.5331	0.5023	0.4990	0.0538	0.0477	0.0452	0.0443
CI^{δ^*}	0.9077	0.9954	0.9994	0.9997	0.9085	0.9956	0.9992	0.9997
CI^{τ^*}	0.9077	0.8752	0.8547	0.8412	0.9085	0.8775	0.8539	0.8408
CI^{η^*}	0.7356	0.7844	0.8187	0.8403	0.3806	0.3836	0.3926	0.3972
$\tilde{\text{CI}}^{\delta^*}$	0.9986	1.0000	1.000	1.0000	0.9987	1.0000	1.0000	1.0000
$\tilde{\text{CI}}^{\tau^*}$	0.9986	0.9989	0.9995	0.9993	0.9987	0.9991	0.9996	0.9992
$\tilde{\text{CI}}^{\eta^*}$	0.9967	0.9985	0.9993	0.9995	0.9860	0.9899	0.9927	0.9940
$\text{R}(\text{CI}^{\delta^*})$	1.1011	1.3918	1.5248	1.5820	1.1036	1.3921	1.5230	1.5780
$\text{R}(\text{CI}^{\tau^*})$	1.1011	0.8668	0.7673	0.7233	1.1036	0.8676	0.7694	0.7271
$\text{R}(\text{CI}^{\eta^*})$	1.0717	0.9491	0.8951	0.8896	0.3067	0.2715	0.2577	0.2522

TABLE 4. Estimation and test performance under the PTA at nominal level of significance $\alpha = 5\%$ conditional on all pre-treatment betas being insignificant at 5% level of significance.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\hat{\beta}_{T+1}$	−0.2505	−0.2547	−0.2503	−0.2465	−0.2506	−0.2513	−0.2516	−0.2511
$\text{CI}^{\hat{\beta}_{T+1}}$	0.8589	0.8496	0.8405	0.8386	0.2687	0.2572	0.2361	0.2367
δ^*	0.8066	1.0036	1.0989	1.1430	0.4136	0.4687	0.5022	0.5161
τ^*	0.8066	0.7226	0.6913	0.6786	0.4136	0.3914	0.3847	0.3818
η^*	0.9263	0.8532	0.8086	0.7900	0.1941	0.1778	0.1744	0.1725
CI^{δ^*}	0.9243	0.9981	0.9998	0.9999	0.9629	0.9976	0.9997	0.9999
CI^{τ^*}	0.9243	0.9331	0.9300	0.9317	0.9629	0.9692	0.9743	0.9732
CI^{η^*}	0.7730	0.8229	0.8338	0.8315	0.2625	0.2032	0.1853	0.1833
$\tilde{\text{CI}}^{\delta^*}$	0.9986	1.0000	1.000	1.0000	0.9995	1.0000	1.0000	1.0000
$\tilde{\text{CI}}^{\tau^*}$	0.9986	0.9997	0.9996	0.9999	0.9995	1.0000	1.0000	1.0000
$\tilde{\text{CI}}^{\eta^*}$	0.9963	0.9994	0.9987	0.9988	0.8828	0.8684	0.8640	0.8578
$\text{R}(\text{CI}^{\delta^*})$	1.3326	1.7079	1.9114	2.0039	2.1817	2.5463	2.7857	2.8867
$\text{R}(\text{CI}^{\tau^*})$	1.3326	1.2296	1.2017	1.1893	2.1817	2.1265	2.1342	2.1355
$\text{R}(\text{CI}^{\eta^*})$	1.5287	1.4503	1.4047	1.3839	1.0238	0.9659	0.9674	0.9651

TABLE 5. 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 $V_i \overset{\text{i.i.d}}{\sim} \text{N}(\frac{1}{4}, 1)$) at nominal level of significance $\alpha = 5\%$ with true ATT $\beta_{T+1} = 0$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\hat{\beta}_{T+1}$	−0.4979	−0.4990	−0.4962	−0.4988	−0.4996	−0.5031	−0.4998	−0.5004
$\mathrm{CI}^{\hat{\beta}_{T+1}}$	0.6287	0.6093	0.5990	0.5827	0.0004	0.0004	0.0008	0.0003
δ^*	1.0203	1.2028	1.3045	1.3538	0.6633	0.7204	0.7501	0.7659
τ^*	1.0203	0.9465	0.9215	0.9115	0.6633	0.6433	0.6327	0.6315
η^*	1.3089	1.1785	1.1329	1.1432	0.4959	0.4696	0.4521	0.4496
CI^{δ^*}	0.9512	0.9958	0.9998	0.9999	0.9637	0.9975	0.9994	0.9997
CI^{τ^*}	0.9512	0.9600	0.9656	0.9623	0.9637	0.9708	0.9768	0.9740
CI^{η^*}	0.8099	0.8435	0.8501	0.8501	0.4180	0.3410	0.2982	0.2856
$\tilde{\mathrm{CI}}^{\delta^*}$	0.9990	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.0000
$\tilde{\mathrm{CI}}^{\tau^*}$	0.9990	1.0000	1.000	0.9999	1.0000	1.0000	1.0000	1.0000
$\tilde{\mathrm{CI}}^{\eta^*}$	0.9952	0.9983	0.9990	0.9988	0.8714	0.8629	0.8517	0.8498
$\mathrm{R}(\mathrm{CI}^{\delta^*})$	1.6834	2.0472	2.2686	2.3746	3.5009	3.9143	4.1603	4.2836
$\mathrm{R}(\mathrm{CI}^{\tau^*})$	1.6834	1.6105	1.6021	1.6031	3.5009	3.4951	3.5090	3.5321
$\mathrm{R}(\mathrm{CI}^{\eta^*})$	2.1568	2.0040	1.9680	2.0033	2.6171	2.5517	2.5074	2.5146

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 $V_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(\frac{1}{2}, 1)$) at nominal level of significance $\alpha = 5\%$ with true ATT $\beta_{T+1} = 0$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.9491	0.8673	0.7088	0.5454	0.9434	0.7839	0.2996	0.0297
$\hat{\beta}_{T+1}$	0.0254	0.0239	0.0250	0.0251	0.0260	0.0255	0.0247	0.0255
$\mathrm{CI}^{\hat{\beta}_{T+1}}$	0.9501	0.9490	0.9487	0.9494	0.9390	0.9398	0.9422	0.9405
δ^*	0.6415	0.8224	0.9420	1.0227	0.2052	0.2762	0.3653	0.4607
τ^*	0.6415	0.5251	0.5046	0.5199	0.2052	0.1807	0.2114	0.2580
η^*	0.5781	0.5175	0.5121	0.5275	0.0598	0.0569	0.0695	0.0935
CI^{δ^*}	0.9070	0.9955	0.9991	0.9994	0.9039	0.9899	0.9978	0.9996
CI^{τ^*}	0.9070	0.8867	0.8741	0.8762	0.9039	0.8727	0.9007	0.9472
CI^{η^*}	0.71486	0.7744	0.8018	0.8214	0.3931	0.4071	0.4996	0.6217
$\tilde{\mathrm{CI}}^{\delta^*}$	0.9986	1.0000	1.000	1.0000	0.9982	1.0000	1.0000	1.0000
$\tilde{\mathrm{CI}}^{\tau^*}$	0.9986	0.9991	0.9988	0.9986	0.9982	0.9979	0.9962	0.9972
$\tilde{\mathrm{CI}}^{\eta^*}$	0.9963	0.9980	0.9986	0.9985	0.9810	0.9816	0.9833	0.9859
$\mathrm{R}(\mathrm{CI}^{\delta^*})$	1.1477	1.4801	1.6969	1.8430	1.1692	1.5742	2.0820	2.6253
$\mathrm{R}(\mathrm{CI}^{\tau^*})$	1.1477	0.9450	0.9090	0.9369	1.1692	1.0297	1.2044	1.4705
$\mathrm{R}(\mathrm{CI}^{\eta^*})$	1.0366	0.9305	0.9219	0.9502	0.3406	0.3241	0.3960	0.5327

TABLE 7. 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 true ATT $\beta_{T+1} = 0$.

	$n_t = 100$				$n_t = 1000$			
	$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 2$	$T = 4$	$T = 8$	$T = 12$
$\#insig/M$	0.9448	0.8433	0.5492	0.2342	0.9148	0.5291	0.0068	0.0000
$\hat{\beta}_{T+1}$	0.0483	0.0508	0.0520	0.0527	0.0504	0.0500	0.0496	0.0499
$\text{CI}^{\hat{\beta}_{T+1}}$	0.9475	0.9469	0.9466	0.9452	0.9132	0.9141	0.9157	0.9145
δ^*	0.6436	0.8388	1.0182	1.1939	0.2148	0.3221	0.5168	0.7172
τ^*	0.6436	0.5400	0.5640	0.6423	0.2148	0.2200	0.3109	0.4088
η^*	0.5866	0.5336	0.5666	0.6590	0.0640	0.0718	0.1281	0.2206
CI^{δ^*}	0.9088	0.9936	0.9979	0.9992	0.8822	0.9823	0.9995	1.0000
CI^{τ^*}	0.9088	0.8793	0.8746	0.8993	0.8822	0.8737	0.9656	0.9945
CI^{η^*}	0.7293	0.7785	0.8210	0.8613	0.3909	0.4657	0.6945	0.8776
$\tilde{\text{CI}}^{\delta^*}$	0.9983	1.0000	1.000	1.0000	0.9948	1.0000	1.0000	1.0000
$\tilde{\text{CI}}^{\tau^*}$	0.9983	0.9985	0.9978	0.9967	0.9948	0.9926	0.9986	0.9998
$\tilde{\text{CI}}^{\eta^*}$	0.9964	0.9976	0.9978	0.9982	0.9651	0.9686	0.9822	0.9951
$\text{R}(\text{CI}^{\delta^*})$	1.1556	1.5091	1.8339	2.1516	1.2235	1.8354	2.9450	4.0875
$\text{R}(\text{CI}^{\tau^*})$	1.1556	0.9714	1.0158	1.1574	1.2235	1.2536	1.7717	2.3297
$\text{R}(\text{CI}^{\eta^*})$	1.0517	0.9591	1.0201	1.1873	0.3643	0.4093	0.7297	1.2570

TABLE 8. 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 true ATT $\beta_{T+1} = 0$.