Direct and Indirect Treatment Effects with Time-Varying Covariates[*]

Nicholas L. Brown
Florida State University

Kyle Butts
University of Arkansas

Joakim Westerlund[†]
Lund University
and
Deakin University

December 9, 2024

Abstract

We propose a simple approach to treatment effect estimation that is valid when the number of time periods is small and the parallel trends condition is violated due to the presence of interactive fixed effects. We show that if there are time-varying covariates that are linear in the same factors as the outcome variable, we can estimate not only the usual dynamic average treatment effects on the treated, but we can also separate the effect of treatment into different causal channels. The asymptotic properties of the estimator are established and their accuracy in small samples is investigated using Monte Carlo simulations. The procedure is illustrated using as an example the effect of increased trade competition on firm markups in China. We estimate that about half of the impact of China's entrance into the WTO on markup dispersion came from the changes in industry-level productivity.

JEL Classification: C31, C33, C38.

Keywords: Treatments effects estimation; fixed-T; interactive fixed effects; common correlated effects; mediation analysis.

[*] Westerlund would like to thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

[†] Corresponding author: Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

1 Introduction

Treatment effects studies using panel data typically estimate linear models with additive unit and time fixed effects. Such fixed effects are consistent with the so-called "parallel trends" assumption, which demands that the average outcomes for treated and untreated cross-sectional units would have followed the same path over time in absence of treatment. Because this assumption is implausible in many settings, much effort has been spent trying to control for violations through the inclusion of covariates. A common approach estimates either time-varying slopes on time-invariant/'baseline' covariates or unit-varying slopes on covariates that only vary over time, like linear time trends (Abadie, 2005; Kim and Oka, 2014; Callaway and Sant'Anna, 2021; Wooldridge, 2021; Borusyak et al., 2024). These methods usually model parallel trend deviations as a product of time-varying slopes and baseline covariates: the former captures the trend itself, and the latter measures the unit's exposure to the trend. The problem is that many drivers of non-parallel trending are unobserved and lack good proxies.[1] Moreover, covariates that vary over time and across individuals likely contain more information on the non-trending behavior. However, when these covariates are affected by treatment, including them in estimation will introduce bias and may cause more harm than good (Angrist and Pischke, 2009; Caetano and Callaway, 2023).

The present paper is motivated by the problem of incorporating data sets with rich controls into treatment effect estimation. We start by modeling the non-parallel trending behavior via interactive fixed effects, which are the product of time effects (called "factors") and unit effects (called "factor loadings").[2] Our approach assumes that researchers have access to a set of time-varying covariates that are impacted by the same set of time effects that drive non-parallel trending behavior in the outcome variable. This assumption coincides with the common correlated effects (CCE) setting of Pesaran (2006), a popular approach in the factor-model literature and one that is known to perform well in finite samples (Westerlund et al., 2019). We then show

[1] For example, in studies with time-varying-only covariates, deterministic linear and quadratic trends are often the only candidates, even though there is plenty of evidence to suggest that they are not enough (Kim and Oka, 2014).

[2] The usual two-way error model with additive time and unit fixed effects is a special case of the interactive effects model. As such, parallel trends is a special case of our modeling assumptions.

that dynamic average treatment effects on the treated are estimable by recovering the unobserved factors via cross-sectional averages of the outcome variable and the covariates.

Our approach is similar in spirit to ?. They consider a model where treatment is correlated with an unobserved confound $c_{i,t}$. They also assume the existence of covariates that are linear in $c_{i,t}$ but unaffected by treatment. These covariates are then used to control for deviations from parallel trends due to $c_{i,t}$. Our approach primarily differs in that we impose separability in $c_{i,t}$: we assume it can be expressed as the product of two unobserved components, $f_t \times \lambda_i$. In doing so, we can recover the untreated potential outcomes post-treatment and allow the covariates to change arbitrarily after receiving treatment. This approach offers some benefits over alternative methods, which we discuss below. Moreover, our treatment effect estimator can allow for arbitrary treatment effect heterogeneity, which has proven a concern for many standard estimators.

Our paper allows for treatment to change the value of the covariates which then have a subsequent impact on the outcome, what the mediation literature calls this an 'indirect effect'. The factor assumption on the covariates allows us to estimate how much treatment shifts the value of the control variables. When the covariates directly affect the outcome, i.e. the causal effect of x on y is non-zero, our estimator can breakdown how much of the estimated treatment effect is driven by treatment changing the value of a specific covariate. Applied work will often perform auxiliary treatment effect estimates using control variables as an outcome to assess if a causal mechanism operates through some covariate. Our method formalizes this kind of analysis to quantify the magnitude of potential mechanisms.

Our proposed treatment effect estimator, dubbed "TECCE", is computed in three simple steps. We begin by estimating the common factors using the CCE approach of forming cross-sectional averages of all the observables from the never-treated sample. We then estimate the factor loadings by regressing the outcome variable for each cross-sectional unit on the first-step factor estimates. In the third and final step, we estimate the counterfactual outcome by taking the product of the first-step factor and second-step loading estimates. Average treatment effects are estimated as the average difference between the observed treated and estimated counterfactual outcomes.

Researchers interested in studying the extent to which the covariates contribute to the ATT can do so by repeating steps two and three of the estimation procedure while conditioning on the observed (possibly treatment-affected) covariates. This procedure gives an estimator of the part of the ATT that is not due to the covariates, which can be subtracted off the initial ATT estimator to produce an estimator of the part of the ATT that is due to the covariates.

The new estimator is shown to be consistent and asymptotically normal under general conditions provided only that the number of treated and untreated units is large, a result that is verified in finite samples by means of Monte Carlo simulations. This result is noteworthy because the TECCE estimation procedure described in the previous paragraph does not require explicit accounting of the covariates in steps two and three. In spite of this, consistency and asymptotic normality hold regardless of whether the covariate values are changed by treatment.

There are three primary benefits to our estimator. First, and most importantly, it allows for treatment-affected covariates and enables researchers to separate the part of the treatment effect that is due to the covariates changing values from the part that is not. We can even obtain valid confidence intervals for the distinct causal mechanisms. As far as we are aware, there is currently no other treatment effects approach with this level of flexibility.

Second, our new estimator does not require the number of time periods T to be large. Interactive fixed effects approaches tend to be "data hungry"; a common requirement is that both T and N are large (Gobillon and Magnac, 2016; Xu, 2017; Arkhangelsky et al., 2021; Chan and Kwok, 2022), which is a problem because in treatment effects studies T is often small (Bertrand et al., 2004). The new procedure accounts for this smallness of T and is as a result widely applicable to applied microeconomic problems.

Finally, it is user-friendly and computationally simple. Estimation only requires the computation of sample averages and linear regression. Gobillon and Magnac (2016), Xu (2017), and Chan and Kwok (2022) use principal components (PC) estimation to control for the factors. Such an approach is based on solving a non-convex optimization problem, which means that it can be difficult to get to converge.[3] Brown and Butts (2023) and Callaway and Karami (2023) use

[3] Even if it does converge, it may not be to the global optimum (Moon and Weidner, 2019).

overidentified generalized method of moments (GMM) estimators, requiring valid instruments that may not be available in practice. Callaway and Tsyawo (2023) use properties of a staggered-rollout setting to create instruments, but are limited in the parameters they can identify. These methods may also not converge to a global maximum in practice (Hayakawa, 2016). Both GMM and PC approaches require knowing the number of unobserved effects, which is a difficult estimation problem (Moon and Weidner, 2019; Breitung and Hansen, 2021). Even if the number of factors is estimable, the resulting PC and GMM estimators still suffer from "post-selection bias" see (Leeb and Pötscher, 2005). The new procedure does not require accurate estimation of the number of factors.

As an empirical illustration, we consider the effect of China's accession into the World Trade Organization (WTO) in 2001 on the dispersion of industry-level markups. Our results suggest that the increased competition generated by the accession lead to reduced markup dispersion. Moreover, we find that almost half of this reduction was brought about by a decrease in marginal cost dispersion.

The rest of the paper is structured as follows. Section 2 presents the model and defines the ATT, the estimation of which is the concern of Section 3. Sections 4, 5 and 6 contain the asymptotic, Monte Carlo results, and empirical studies, respectively. Section 7 concludes. All proofs are relegated to the appendix.

2 The model

We are interested in estimating the ATT of a particular treatment on some outcome variable $y_{i,t}$, observable for $i \in \{1, ..., N\}$ cross-sectional units and $t \in \{1, ..., T\}$ time periods. We allow for the possibility that the N units can be divided into groups within which treatment timing is the same. There are $G < \infty$ such groups indexed by $g \in \{1, ..., G\}$. Treated units never leave their groups but remain exposed for all periods after entering treatment; that is, treatment is of the "staggered roll-out" type. Untreated units are members of group $g = 0$.

The timing of the treatment could be viewed as the outcome of a random process that is dealt with through suitable conditioning. We instead follow Arkhangelsky et al. (2021) and Borusyak et al. (2024) and view both the treatment timing and the drivers of non-parallel trending as fixed, which has the advantage that it places no assumption on their joint distribution or the distribution of treatment timing.[4] It is particularly well suited for applications where treatment timing is not explicitly randomized, as in our empirical illustration of Section 6. We denote by $g_i \in \{0, ..., G\}$ a variable stating the group membership of cross-sectional unit i, and by $\mathcal{I}_g := \{i : g_i = g\} \subset \{1, ..., N\}$ the set of cross-sectional units that are members of group $g \in \{0, ..., G\}$, where $a := b$ means that a is defined by b. The set of untreated units is denoted $\mathcal{I}_0$, and it is convenient to let $\mathcal{I}_0^c := \{1, ..., N\} \setminus \mathcal{I}_0$ denote the set of treated units. The number of cross-sectional units within group g is given by the cardinality $N_g := |\mathcal{I}_g|$. Exact conditions on $N_0, ..., N_G$ will be specified later. For now, we just assume that these quantities are "large".[5] We denote T_0 as the first period before any unit is treated, so that the first treatment starts in period $T_0 + 1$. Unlike $N_0, ..., N_G$, the number of pre-treatment periods T_0 does not have to be large, provided that it is larger than the number of estimated factors. We describe this condition later. The start of the treatment of group $g > 0$ is denoted t_g.

Denote by $y_{i,t}(g)$ the "potential" outcome of cross-sectional unit i had it been member of group g in period t. The observed outcome for unit i at time t is given by $y_{i,t} := y_{i,t}(g_i)$. In this notation, the unit-specific treatment effect for a unit i that is member of treatment group $g_i = g > 0$ in post-treatment periods $t > T_0$ is given by

$$\Delta_{i,t} := y_{i,t} - y_{i,t}(0) \tag{1}$$

where $y_{i,t}(0)$ is the untreated potential outcome.[6] Because we do not observe $y_{i,t}(0)$ for treated units in post-treatment periods, $\Delta_{i,t}$ must be treated as unknown and estimated from the data. Of course, because T is fixed, accurate estimation of $\Delta_{i,t}$ itself is impossible. The object of interest is therefore not $\Delta_{i,t}$ but the ATT, $\mathbb{E}(\Delta_{i,t})$, which we now characterize.

Assumption 1. $\Delta_{i,t} = \text{ATT}_{g,t} + v_{i,t}$, where $\text{ATT}_{g,t}$ is non-random and $v_{i,t}$ is a mean zero random

[4] In the below notation, expectations are taken over particular treatment cohorts.

[5] Hence, as usual, we assume that there are both treated and untreated units in the sample. The condition that $N_0, ..., N_G$ are all large is also very common, although it is not always articulated in the same way as here. Many studies assume the probability of treatment is strictly positive (Abadie, 2005; Callaway and Sant'Anna, 2021; Sant'Anna and Zhao, 2020), which in the current fixed treatment allocation context is tantamount to requiring that $N_g/N \to \delta > 0$ as N_g, $N \to \infty$.

[6] See Borusyak et al., 2024 for a similar definition

error that is independent of all other random elements of the model.

Assumption 1 implies that $\mathbb{E}(\Delta_{i,t}) = \text{ATT}_{g,t}$, which is the same "group-time" ATT considered by Callaway and Sant'Anna (2021).[7] Except for the constancy-within-groups condition, the ATT is unrestricted, which means that it is allowed to vary freely over both groups and time. One implication of this is that the effect of the treatment need not take place abruptly but can be gradual in nature. This is in contrast to existing studies that typically make quite strong assumptions about the variability of the ATT.

Assumption 2. $y_{i,t} = y_{i,t}(0)$ for $t \leq T_0$.

Assumption 2 is similar to Assumption 11 in Callaway and Karami (2023) and Assumption 3 in Callaway and Sant'Anna (2021). It requires that there is no treatment effect before any unit receives treatment. Once one group enters treatment, however, anticipation is not ruled out for other groups. This condition seems reasonable because before the first group enters treatment, the units might not know what to expect. Once treatment roll-out has commenced, however, it is possible that not-yet-treated units learn from the treated.

We typically do not observe $y_{i,t}$ in isolation but together with covariates whose outcome may again depend on treatment status. In our empirical illustration, we estimate the effect of China's accession into the WTO on industry-level markup dispersion while controlling for the dispersion in marginal costs. According to theory, increased competition may affect markups directly through changes in market structure and indirectly through changes in marginal costs. We therefore introduce the $m \times 1$ vector $\mathbf{x}_{i,t}(g)$, whose realized value is given by $\mathbf{x}_{i,t} := \mathbf{x}_{i,t}(g_i)$.[8]

Assumption 3.

$$y_{i,t}(0) = \boldsymbol{\beta}'\mathbf{x}_{i,t}(0) + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t},$$

[7] Most studies in the literature do not use random coefficient conditions like Assumption 1, but state their conditions directly in terms of the (conditional) expectations of those random coefficients (Gobillon and Magnac, 2016; Callaway and Sant'Anna, 2021; Chan and Kwok, 2022).

[8] Note how $\mathbf{x}_{i,t} = \mathbf{x}_{i,t}(0)$ for $t \leq T_0$ is a necessary condition for Assumption 2. We therefore require that $\mathbf{x}_{i,t}$ satisfies the same anticipation condition as $y_{i,t}$.

where $\boldsymbol{\beta}$ is a $m \times 1$ vector of slope coefficients, $\mathbf{f}_t$ is a $r \times 1$ vector of unobservable common factors, $\boldsymbol{\alpha}_i$ is a $r \times 1$ vector of factor loadings, and $\varepsilon_{i,t}$ is a mean zero random error. Both $\mathbf{f}_t$ and $\boldsymbol{\alpha}_i$ are assumed to be non-random.

Assumption 3 specifies $y_{i,t}(0)$ as a linear function of $\mathbf{x}_{i,t}(0)$, which is commonly imposed in empirical and econometric work, including Wooldridge (2021), Callaway and Karami (2023), Callaway and Tsyawo (2023), Brown and Butts (2023), and Borusyak et al. (2024).[9] We differ from these approaches in that we allow the covariates' distributions to vary with treatment status.

The interactive fixed effects are given by $\boldsymbol{\alpha}_i'\mathbf{f}_t$. The standard approach in the treatment effects literature is to include additive time and unit fixed effects (see Caetano and Callaway, 2023, and Callaway and Karami, 2023 for recent discussions). Such effects are nested within our interactive specification, as is clear from noting that $\boldsymbol{\alpha}_i'\mathbf{f}_t = \eta_i + \theta_t$ in the special case when $\mathbf{f}_t = [1, \theta_t]'$ and $\boldsymbol{\alpha}_i = [\eta_i, 1]'$. A major advantage of the interactive specification when compared to the additive one is that it accommodates violations of the parallel trends condition. Indeed, unless $\mathbf{f}_t = \mathbf{f}$ for all t, trends will not be parallel unless $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ for all pairs i and j.[10] Hence, by leaving $\mathbf{f}_t$ and $\boldsymbol{\alpha}_i$ unrestricted, we can accommodate very general forms of non-parallel behavior. In fact, the interactive effects specification considered here is general even when compared to other specifications of the same type, which typically assume that either $\mathbf{f}_t$ or $\boldsymbol{\alpha}_i$, or both, follow certain probability laws (Gobillon and Magnac, 2016; Xu, 2017; Callaway and Karami, 2023; Callaway and Tsyawo, 2023; Brown and Butts, 2023).[11] Because the factor model and treatment timing is fixed, they are allowed to be arbitrarily related to each other as in Arkhangelsky et al. (2021)

[9] There are some exceptions based on non-parametric approaches (Abadie, 2005; Sant'Anna and Zhao, 2020; Callaway and Sant'Anna, 2021). However, these suffer from the "curse of dimensionality" problem, the solution of which typically involves imposing additional structure, and still the small-sample properties can be unacceptably poor unless sample sizes are very large, so linearity is often used in applications. Non-parametric approaches typically also place strong distributional assumptions directly on the observed data. A very common condition is that $y_{i,t}$ and $\mathbf{x}_{i,t}$ are independently and identically distributed (iid) over i, which is unrealistic.

[10] The parallel trends condition requires that $\mathbb{E}[\Delta y_{i,t}(0)|\Delta \mathbf{x}_{i,t}(0)] = \mathbb{E}[\Delta y_{j,t}(0)|\Delta \mathbf{x}_{j,t}(0)]$ for all $i \in \mathcal{I}_0$, $j \in \mathcal{I}_0^c$ and $t > T_0$, which in terms of the model in Assumption 1 reads $\boldsymbol{\beta}'\Delta\mathbf{x}_{i,t}(0) + \boldsymbol{\alpha}_i'\Delta\mathbf{f}_t = \boldsymbol{\beta}'\Delta\mathbf{x}_{j,t}(0) + \boldsymbol{\alpha}_j'\Delta\mathbf{f}_t$. A necessary condition for this last equality to hold is that the factor loadings are equal on average between treated and never-treated groups (Brown and Butts, 2023).

[11] An example of a common assumption is that $\mathbf{f}_t$ is generated from a stationary stochastic process like in Gobillon and Magnac (2016) and Xu (2017), which is likely restrictive in the present context because it rules out factors that are, for example, breaking or trending.

and Borusyak et al. (2024).

Assumption 4.

$$\mathbf{x}_{i,t}(0) = \boldsymbol{\lambda}_i' \mathbf{f}_t + \mathbf{v}_{i,t},$$

where $\boldsymbol{\lambda}_i$ is a $r \times m$ matrix of non-random factor loadings and $\mathbf{v}_{i,t}$ is a $m \times 1$ vector of mean zero errors that are independent of $\varepsilon_{i,t}$.

Assumption 4 allow $\mathbf{x}_{i,t}(0)$ to load on the same set of factors as $y_{i,t}(0)$, which means that it may be endogenous. Hence, in contrast to much of the previous literature, here treatment is not the only source of endogeneity. The common dependence on $\mathbf{f}_t$ is consistent with the empirical observation that many variables are affected by the same common shocks (Westerlund et al., 2019). The factors need not be the same, though, as loadings may be zero. There might therefore be factors that are unique to $\mathbf{x}_{i,t}(0)$ and/or $y_{i,t}(0)$.

The condition that $\mathbf{x}_{i,t}(0)$ loads linearly on $\mathbf{f}_t$ rules out specifications featuring, for example, dummy variables, powers or products of the covariates. However, additional covariates can be easily accommodated in the outcome model provided that they satisfy standard exogeneity conditions, as we explain in the online appendix.[12] These covariates can be 'controlled for' in the final imputation procedure, but are not useful for estimating $\mathbf{f}_t$, and so we cannot use them to estimate a causal channel.

Most papers in the literature do not place any parametric assumptions on the covariates.[13] However, they assume instead that the covariates are exogenous (to treatment) and iid over the cross-section (Abadie, 2005; Callaway and Sant'Anna, 2021; Callaway and Karami, 2023; Brown and Butts, 2023; Callaway and Tsyawo, 2023; Caetano and Callaway, 2023), which is more restrictive than Assumption 4. In fact, the only studies that come close to ours in terms of the generality of the allowable covariates are Gobillon and Magnac (2016), Xu (2017), and Chan and Kwok (2022). They allow the covariates to be arbitrarily correlated with the interactive

[12] If there are no covariates available, we define $\boldsymbol{\beta}' \mathbf{x}_{i,t}(0) := 0$ in Assumption 3. In this case, Assumption 4 is not needed. Our estimator can still be constructed using only the outcome variable.

[13] One exception is given by Caetano and Callaway (2023), which is also one of the few studies in the literature to allow for treatment-affected covariates.

fixed effects, which represents a more general consideration than Assumption 4. However, they require instead that the covariates are stationary and unaffected by treatment, which is not required here.[14]

Caetano and Callaway (2023) is one of the few papers in the literature to allow for covariates affected by treatment. They focus on the ATT; however, they also say that "it would be interesting to extend our arguments to additionally identifying direct and indirect effects of participating in the treatment" (page 6). We provide such an extension. In order to separate the part of the ATT that is due to changing values of the covariates from the part that is not, we define

$$\boldsymbol{\tau}_{i,t} := \mathbf{x}_{i,t} - \mathbf{x}_{i,t}(0) \tag{2}$$

for a unit i that is member of group $g_i = g > 0$ in post-treatment periods $t > T_0$. In view of Assumption 3, the covariates' contribution to the unit-specific treatment effect on $y_{i,t}$ is given by $\boldsymbol{\beta}'[\mathbf{x}_{i,t}(g) - \mathbf{x}_{i,t}(0)] = \boldsymbol{\beta}'\boldsymbol{\tau}_{i,t}$. We can allow $\boldsymbol{\tau}_{i,t} = 0$ so that the covariates are not affected by treatment. Thus, we can accommodate covariates that are linear in the common factors, whether or not they are affected by treatment, along with covariates that are not affected by treatment but may have a nonlinear relationship to the factors. We also define the difference between these effects;

$$\eta_{i,t} := \Delta_{i,t} - \boldsymbol{\beta}'\boldsymbol{\tau}_{i,t}. \tag{3}$$

In the terminology of the mediation literature, $\eta_{i,t}$ is the "direct" effect of treatment and $\boldsymbol{\beta}'\boldsymbol{\tau}_{i,t}$ is the "indirect" effect of treatment mediated through the covariates (Huber, 2014).

The random coefficient condition in Assumption 1 is enough if the purpose is to just estimate the ATT. If the purpose is to estimate also the direct and indirect ATTs, henceforth abbreviated "DATT" and "IATT", respectively, more conditions of the same type are needed.

Assumption 5. $\boldsymbol{\tau}_{i,t} = \boldsymbol{\tau}_{g,t} + \boldsymbol{\zeta}_{i,t}$, where $\boldsymbol{\tau}_{g,t}$ is non-random, and $\boldsymbol{\zeta}_{i,t}$ are mean zero random errors that are independent of all other random elements of the model.

[14] They also require that T is large and that the number of common factors, r, can be accurately estimated, which is again not a requirement in the present paper.

Assumption 5 implies that

$$\mathbb{E}(\boldsymbol{\beta}'\boldsymbol{\tau}_{i,t}) = \boldsymbol{\beta}'\boldsymbol{\tau}_{g,t} =: \text{IATT}_{g,t}. \tag{4}$$

Further use of Assumption 1 gives

$$\mathbb{E}(\eta_{i,t}) = \mathbb{E}(\Delta_{i,t}) - \mathbb{E}(\boldsymbol{\beta}'\boldsymbol{\tau}_{i,t}) = \text{ATT}_{g,t} - \text{IATT}_{g,t} =: \text{DATT}_{g,t} \tag{5}$$

or, equivalently,

$$\text{ATT}_{g,t} = \text{DATT}_{g,t} + \text{IATT}_{g,t}. \tag{6}$$

3 The TECCE estimator

3.1 The ATT

We need an estimate of $y_{i,t}(0)$ in order to estimate the ATT. Interestingly, because the objective here is to estimate only the ATT and not the other parameters of the model, the estimation of $y_{i,t}(0)$ only requires accounting for the interactive fixed effects. We use the fact that the never-treated data is asymptotically linear in the common factors. For example, note that

$$\frac{1}{N_0}\sum_{i\in\mathcal{I}_0} \mathbf{x}_{i,t} = \overline{\boldsymbol{\lambda}}'\mathbf{f}_t + o_p(1) \tag{7}$$

for $t = 1, ..., T$. We can also include the outcome's average because Assumptions 3 and 4 imply that $y_{i,t}$ is also linear in $\mathbf{f}_t$. Thus, the cross-sectional averages of the data can be used to asymptotically control for the factor space and allow us to recover the untreated potential outcomes for treated units.[15]

Counterfactual estimation procedure:

1. For all t, compute

$$\widehat{\mathbf{f}}_t := \frac{1}{N_0}\sum_{i\in\mathcal{I}_0} \mathbf{z}_{i,t}, \tag{8}$$

where $\mathbf{z}_{i,t} := [y_{i,t}, \mathbf{x}'_{i,t}]'$ is a $(m+1)\times 1$ vector containing all the observables.

[15] The cross-sectional averages are not consistent estimates of the factors because the factor loadings are also unobserved. However, they will asymptotically span the space that contains the common factors, which is sufficient for our purposes. See Westerlund et al. (2019) for details.

2. Estimate the following regression by OLS for all i and $t \leq T_0$:

$$y_{i,t} = \mathbf{a}_i'\widehat{\mathbf{f}}_t + u_{i,t}, \tag{9}$$

where $\mathbf{a}_i$ is a $(m+1) \times 1$ vector of factor loadings and $u_{i,t}$ is an error term. Define the $T_0 \times 1$ vector $\mathbf{y}_i := [y_{i,1}, ..., y_{i,T_0}]'$ and the $T_0 \times (m+1)$ matrix $\widehat{\mathbf{f}} := [\widehat{\mathbf{f}}_1, ..., \widehat{\mathbf{f}}_{T_0}]'$. In this notation, the OLS estimator of $\mathbf{a}_i$ is given by

$$\widehat{\mathbf{a}}_i := (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'\mathbf{y}_i, \tag{10}$$

which is computed for all i.

3. The sought counterfactual estimator is given by

$$\widehat{y}_{i,t}(0) := \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t, \tag{11}$$

which is available for all treated observations $i \in \mathcal{I}_0^c$ and $t > T_0$. Here $\{\widehat{\mathbf{a}}_i\}_{i \in \mathcal{I}_0^c}$ is from step 2, while $\{\widehat{\mathbf{f}}_t\}_{t > T_0}$ is from step 1.

The fact that $\widehat{\mathbf{f}}_t$ is computed with the untreated units only is crucial since in the present paper both $y_{i,t}$ and $\mathbf{x}_{i,t}$ can depend on the treatment, which may lead to inconsistency if we use the entire sample. With $y_{i,t}(g) = y_{i,t}$ known and $y_{i,t}(0)$ estimated, the estimated treatment effect is given by

$$\widehat{\Delta}_{i,t} := y_{i,t} - \widehat{y}_{i,t}(0) \tag{12}$$

for $i \in \mathcal{I}_0^c$ and $t > T_0$. The TECCE estimator of $\mathrm{ATT}_{g,t}$ for group $g > 0$ at time t is obtained by averaging over all group members;

$$\widehat{\mathrm{ATT}}_{g,t} := \frac{1}{N_g} \sum_{i \in \mathcal{I}_g} \widehat{\Delta}_{i,t}. \tag{13}$$

A major point about the above estimation procedure is that there is no need to account for the covariates in steps 1 and 2. The intuition for why is as follows: the estimated unit-specific treatment effect can be written as

$$\widehat{\Delta}_{i,t} = y_{i,t} - \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t = y_{i,t} - \mathbf{y}_i'\widehat{\mathbf{f}}(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}_t, \tag{14}$$

which is a "defactored" version $y_{i,t}$.[16] Asymptotically the defactoring eliminates the interactive fixed effects. This means that the source of the endogeneity of $\mathbf{x}_{i,t}$ is gone and therefore the covariates can be ignored without risking omitted variables bias. The defactoring is also the reason for why the procedure works in spite of the fact that $\widehat{\mathbf{a}}_i$ is not consistent and in fact remains random even asymptotically because T is fixed. The defactoring leads to increased variance; however, the interactive fixed effects are gone and so the asymptotic validity of the procedure is unaffected. Hence, similarly to studies such as Sant'Anna and Zhao (2020), our main concern here is not efficiency, but robust estimation and inference. If there are covariates present that are either known not to have a factor structure or there is uncertainty over the process that generated them, the above procedure has to be modified as explained in the online appendix.

As Caetano and Callaway (2023) point out, the validity of estimates of the ATT typically depend on whether or not the covariates are affected by treatment status. Most studies circumvent this problem by assuming that there are no covariates at all (Arkhangelsky et al., 2021; Brown and Butts, 2023), that any covariates are time-invariant (Wooldridge, 2021; Callaway and Karami, 2023; Callaway and Tsyawo, 2023) or, perhaps most commonly, that the covariates may depend on time but that they are unaffected by treatment (Gobillon and Magnac, 2016; Xu, 2017; Chan and Kwok, 2022). The problem here is that if there are time-varying covariates present that depend on the treatment status, approaches that fail to properly control for these will be subject to omitted variables bias. However, the solution is not as simple as just including the relevant treatment-affected covariates into the model. For example, if we are estimating the effect of a certain policy aimed at reducing unemployment, we might want to control for the poverty rate.[17] Because such policies might indirectly reduce poverty, the poverty rate covariate will absorb some of the treatment effect, typically referred to as "post-treatment bias". Because of this, treatment-affected covariates are often considered to be "bad controls" (Angrist and Pischke, 2009). There is therefore a dilemma; while including the covariates induces post-treatment bias, excluding them induces omitted variables bias (Aklin and Bayer, 2017). The defactoring

[16] This insight is similar to equations (8) and (9) of ?.

[17] See Caetano and Callaway (2023) for more examples of this type.

implicit in TECCE not only eliminates the source of the omitted variables bias, but also enables estimation of the ATT without including the treatment-affected covariates. It therefore resolves the dilemma.

3.2 The direct and indirect ATTs

We demonstrated in Section 2 that the ATT can be decomposed into the DATT and the IATT. We now show how to estimate these constituent parts. The idea behind our estimator of the DATT builds on the discussion of the previous subsection about bad controls. In particular, by making the counterfactual estimation procedure conditional on observed $\mathbf{x}_{i,t}$, we no longer estimate the ATT but instead estimate the DATT. In other words, including the bad controls induces post-treatment bias that is exactly equal to the DATT.[18]

Controlling for $\mathbf{x}_{i,t}$ requires two changes to the counterfactual estimation procedure presented in Section 3.1. First, the appropriate step-2 regression model to be estimated is no longer given by (9) but by

$$y_{i,t} = \boldsymbol{\beta}'\mathbf{x}_{i,t} + \mathbf{a}_i'\widehat{\mathbf{f}}_t + u_{i,t}. \tag{15}$$

Define the $T_0 \times m$ matrix $\mathbf{x}_i := [\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,T_0}]'$ and the $T_0 \times T_0$ matrix $\mathbf{M_A} := \mathbf{I}_{T_0} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, where $\mathbf{A}$ is any T_0-rowed matrix. The estimators of $\boldsymbol{\beta}$ and $\mathbf{a}_i$ in the above model can now be written in the following way:

$$\widehat{\boldsymbol{\beta}} := \left(\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{x}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{y}_i, \tag{16}$$

$$\widehat{\mathbf{a}}_i := (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}). \tag{17}$$

Second, the step-3 counterfactual estimator is now given by

$$\widehat{y}_{i,t}(0) := \widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,t} + \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t. \tag{18}$$

[18] The fact that controlling the covariates alters the object being estimated is important not only for the present paper, but also when considering the works of others. As alluded to earlier, many studies assume that the covariates are unaffected by treatment and use the observed covariates in their ATT estimations. Logic based on our findings suggests that if the unaffected covariates assumption is false, these estimators will only capture the DATT. The Monte Carlo and empirical studies of Sections 5 and 6 elaborate on this point.

Given the above changes, the estimated DATT for group g at time t is entirely analogous to the estimated ATT;

$$\widehat{\mathrm{DATT}}_{g,t} := \frac{1}{N_g} \sum_{i \in \mathcal{I}_g} \widehat{\eta}_{i,t}, \tag{19}$$

where $\widehat{\eta}_{i,t} := y_{i,t} - \widehat{y}_{i,t}(0)$. The IATT can be estimated by simply subtracting off $\widehat{\mathrm{DATT}}_{g,t}$ from $\widehat{\mathrm{ATT}}_{g,t}$;

$$\widehat{\mathrm{IATT}}_{g,t} := \widehat{\mathrm{ATT}}_{g,t} - \widehat{\mathrm{DATT}}_{g,t}. \tag{20}$$

4 Asymptotic results

We can now study the asymptotic properties of the estimated ATT and its direct and indirect parts. Assumptions 1–5 characterize the model. In order to establish the required asymptotic results, more conditions are need.

Assumption 6. $v_{i,t}$, $\boldsymbol{\zeta}_{i,t}$, $\varepsilon_{i,t}$ and $\mathbf{v}_{i,t}$ are all independently distributed across i with finite fourth-order cumulants. Also, $\mathbf{lim}_{N_g \to \infty} N_g^{-1} \sum_{i \in \mathcal{I}_g} \mathbb{E}(\varepsilon_{i,t}^2) =: \sigma_{\varepsilon,g,t}^2$ and the $m \times m$ matrix $\mathbf{lim}_{N_g \to \infty} N_g^{-1} \sum_{i \in \mathcal{I}_g} \mathbb{E}(\mathbf{v}_{i,t}\mathbf{v}$
$\boldsymbol{\Sigma}_{\mathbf{v},g,t}$ exist and are positive definite for all g.

Assumption 7. The $r \times (m+1)$ matrix $\boldsymbol{\Lambda}_i := [\boldsymbol{\alpha}_i + \boldsymbol{\lambda}_i \boldsymbol{\beta}, \boldsymbol{\lambda}_i]$ is such that $N_0^{-1} \sum_{i \in \mathcal{I}_0} \boldsymbol{\Lambda}_i$ has full row rank r.

Assumption 8. $T_0 > m + 1$.

Assumption 9. The $r \times r$ matrix $\sum_{t=1}^{T_0} \mathbf{f}_t \mathbf{f}_t'$ is positive definite.

Assumption 10. The $m \times m$ matrix $\mathbf{plim}_{N \to \infty} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{M}_{\widehat{\mathbf{f}}} \mathbf{x}_i$ exist and is positive definite.

Assumption 6 is not particularly restrictive in the sense that $y_{i,t}$ and $\mathbf{x}_{i,t}$ are still allowed to be strongly cross-sectionally correlated through $\boldsymbol{\lambda}_i' \mathbf{f}_t$.[19] While we do require that the covariance

[19] We illustrated this point using $\mathbf{x}_{i,t}(0)$. By Assumptions 4 and 6, $\mathbb{E}[\mathbf{x}_{i,t}(0)\mathbf{x}_{j,t}(0)'] = \boldsymbol{\lambda}_i' \mathbf{f}_t \mathbf{f}_t' \boldsymbol{\lambda}_j$, which is not zero unless the loadings are. Interactive fixed effects therefore provide a means to accommodate strong cross-sectional dependence. See Chudik et al. (2011) for a thorough discussion of the notions of weak and strong cross-sectional dependence.

matrices of $\varepsilon_{i,t}$ and $\mathbf{v}_{i,t}$ are positive definite, we place no such restrictions on the covariance matrices of $v_{i,t}$ and $\boldsymbol{\zeta}_{i,t}$.

Step 1 of the counterfactual estimation procedure uses the cross-sectional averages of the observables to estimate the factors. We therefore require that those averages are informative about the factors. Assumption 7 rules out situations in which there are factors present but their effect on $\widehat{\mathbf{f}}_t$ averages out. Another situation that is ruled out by this assumption is when $m+1 \geq r$, so that the number of factors is larger than the number of cross-sectional averages used to estimate them.[20] One way to relax this condition is if some of the factors are "observed", like a unit-specific intercept or time trend, because Assumption 7 only applies to the loadings of unobserved factors. Observed factors can be appended to $\mathbf{f}_t$ in Assumptions 3 and 4, and to $\widehat{\mathbf{f}}_t$ in step 1 of the counterfactual estimation procedure. The rest is unaffected. In Section 6, we elaborate on this point and explain how to test Assumption 7.

Assumptions 8–10 are non-collinearity conditions. Assumptions 8 and 9 ensure that the $(m+1) \times (m+1)$ matrix $\widehat{\mathbf{f}}'\widehat{\mathbf{f}}$ appearing in the step-2 estimator of $\mathbf{a}_i$ is positive definite both asymptotically and in small samples.[21] Assumption 10 generalizes the usual "within assumption" in the unit-specific fixed effects only model, which rules out time-invariant covariates. Assumption 10 rules out more general "low-rank" covariates, such as deterministic constant and trend terms that only vary over time, as it is almost always done in models with interactive effects.[22] The reason for this condition is that if there are such low rank covariates present, they will be captured by $\widehat{\mathbf{f}}_t$, and therefore they cannot be included also in $\mathbf{x}_i$. Fortunately, there is an easy fix to this problem, which is to treat all low rank covariates as observed factors and to append them to $\widehat{\mathbf{f}}$. An important point about Assumptions 6–10 is that the time series properties of $\mathbf{f}_t$, $\varepsilon_{i,t}$ and $\mathbf{v}_{i,t}$ are essentially unrestricted. Note in particular how, unlike most other treatment effects studies, stationarity and homoskedasticity are not needed.

We are now ready to state Theorem 1, which contains our first main result.

[20] See Juodis et al. (2021) for a discussion and some results for the case when Assumption 7 fails.

[21] Because our estimator of $\mathbf{a}_i$ comes from a regression on pre-treatment observations, we need at least as many degrees of freedom as allowable factors. This assumption is similar in principle to requiring at least two time periods for a fixed effects regression.

[22] See Moon and Weidner (2015) for a discussion.

Theorem 1. Suppose that Assumptions 1–4 and 6–9 are met. Then, uniformly in $g \in \{1, ..., G\}$ and $t \in \{T_0 + 1, ..., T\}$, as N_g, $N_0 \to \infty$ with $N_g/N_0 \to \delta < \infty$,

$$\frac{\sqrt{N_g}(\widehat{\text{ATT}}_{g,t} - \text{ATT}_{g,t})}{\sqrt{\text{var}(\widehat{\text{ATT}}_{g,t})}} \to_d N(0,1), \tag{21}$$

where $\to_d$ signifies convergence in distribution and the definition of $\text{var}(\widehat{\text{ATT}}_{g,t})$ is provided in the online appendix.

Theorem 1 does not require that Assumptions 5 and 10 hold, which is natural since the estimation of the ATT does not require explicit conditioning on $\mathbf{x}_{i,t}$. Asymptotic normality therefore holds with only minimal conditions on $\mathbf{x}_{i,t}$.

Theorem 1 implies that $\widehat{\text{ATT}}_{g,t}$ is consistent for $\text{ATT}_{g,t}$ and that the rate of convergence is $N^{-1/2}$. In the treatment effects literature it is common to plot estimates of the ATT over time and to interpret variations as being due to treatment effect dynamics. Theorem 1 says that $\widehat{\text{ATT}}_{g,t}$ can be used for the same purpose. If one is not interested in dynamics but just want a summary measure of the evidence, $\widehat{\text{ATT}}_{g,t}$ can be averaged over g or t.[23] Since G and $T - T_0$ are both fixed, the consistency in Theorem 1 naturally carries over to such averages.

Inference based on Theorem 1 requires a consistent estimator of $\text{var}(\widehat{\text{ATT}}_{g,t})$. Unrestricted treatment effect heterogeneity is known to be problematic for variance estimation in the sense that exact asymptotic inference is not always possible (de Chaisemartin and D'Haultfœuille, 2020; Borusyak et al., 2024). Another problem is the presence of unobserved common factors and the additional estimation uncertainty that they bring. This problem does not necessarily interfere with asymptotic inference; however, it does require additional conditions to ensure that the factor estimation error is negligible. We provide a nonparametric variance estimator in the online appendix that is similar to Pesaran (2006). One can also use the usual nonparametric panel bootstrap.[24]

Theorem 2. Suppose that Assumptions 1–10 are met. Then, the results reported in Theorems 1 and 2 for $\widehat{\text{ATT}}_{g,t}$ apply also to $\widehat{\text{DATT}}_{g,t}$ and $\widehat{\text{IATT}}_{g,t}$.

[23] See Callaway and Sant'Anna (2021) for a thorough discussion of various ways in which $\text{ATT}_{g,t}$ can be meaningfully averaged.

[24] See Westerlund et al. (2019) and Brown et al. (2022).

Theorem 2 implies that the above discussion of Theorems 1 for $\widehat{\text{ATT}}_{g,t}$ applies also to $\widehat{\text{DATT}}_{g,t}$ and $\widehat{\text{IATT}}_{g,t}$. The variance estimates are formed the same way as the $\widehat{\text{ATT}}_{g,t}$ and definitions are given in the online appendix.

5 Simulations

We now present our simulation exercise. We compare our TECCE estimator to two popular estimators that can also accommodate time-varying covariates. The first is the generalized synthetic control (GSC) estimator of Xu (2017). Like our approach, it imputes post-treatment untreated counterfactual outcomes under the assumption of a factor model. It specifically uses the principal components method of Bai (2009) estimated via the untreated observations. It can jointly estimate the unobserved factors, factor loadings, and slope coefficients, but consistency requires $T \to \infty$, so we should expect it to perform poorly when N dominates T. The GSC estimator here uses $r = 1$, which is a courtesy because PC estimation generally requires correct specification of the number of factors, unlike TECCE. We also consider the TWFE imputation estimator of Wooldridge (2021) and Borusyak et al. (2024). They estimate a two-way error model via OLS on the untreated sample and can also include known covariates. We should expect this estimator to perform poorly when parallel trends fails and the unobserved effects take a multiplicative form. For both estimators, we include a specification with and without controlling for covariates linearly.

We generate our data according to the following model:

$$y_{i,t} = d_{i,t} + \mathbf{x}_{i,t} + f_t\alpha_i + \epsilon_{i,t} \tag{22}$$

$$\mathbf{x}_{i,t} = d_{i,t}\tau + f_t\lambda_i + \mathbf{v}_{i,t} \tag{23}$$

where $d_{i,t}$ is an indicator variable that is 1 if unit i is treated at time t. The direct effect η and slope coefficient β are therefore always set to 1. Half of the sample is treated and half is untreated, with treatment occurring in the last time period in each experiment. We generate the factor loadings as

$$\begin{pmatrix} \alpha_i \\ \lambda_i \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 2+\kappa D_i \\ 2+\kappa D_i \end{bmatrix}, \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{bmatrix} \right) \tag{24}$$

The κ parameter indicates a "break" in the factor loadings. When $\kappa = 0$, parallel trends holds because there is no systemic difference in the distribution of the heterogeneity between treated and untreated units. TWFE will be consistent and should perform well in this setting (Callaway and Karami, 2023). We also include an indirect treatment effect τ which we set to different values. When $\tau = 0$, the covariates may help alleviate parallel trend violations because they do not themselves affect the outcome differentially after treatment. The idiosyncratic errors $\epsilon_{i,t}$ and $v_{i,t}$ are iid draws from $N(0, .4^2)$. Adding serial correlation and heteroskedasticity would only further worsen the performance of the OLS estimators GSC and TWFE relative to our TECCE.

We generate two main types of trending behavior. In one set, $f_t = 1$ and trends are parallel, meaning we would expect each estimator to perform reasonably well. In the second set, $f_t = 1 + t/8$, corresponding to a linear trend. This factor model combined with $\kappa \neq 0$ implies a particular type of non-parallel trending behavior. We vary $N \in \{50, 300\}$ and $T \in \{5, 10, 15\}$. For each simulation, we draw f with 15 elements and keep the last T elements so that f_T has the same expected value across simulations.[25] Each experiment is conducted $1,500$ times. Tables 1–3 report the bias and RMSE of each estimator with respect to the overall ATT, which $\Delta = 1$ when $\tau = 0$ and $\Delta = 2$ when $\tau = 1$.

Insert Tables 1, 2, and 3 here

As expected, all three estimators perform well when the factor is constant over time, i.e. parallel trends holds. The only time GSC and TWFE perform poorly under parallel trends is when they control for covariates in the presence of indirect effects. As we described in Section 3, including covariates linearly will absorb the indirect effect of treatment. In this case, the GSC and TWFE estimators controlling linearly for $\mathbf{x}_{i,t}$ have a bias of -1, meaning they are unbiased for the direct effect of treatment, but biased for the overall ATT. We can also see that both TECCE and the two TWFE specifications perform well under the linear trend when $\kappa = 0$. Even though the interactive effect is non-trivial, there are no systemic differences between the treated and untreated groups, so parallel trends holds unconditionally. GSC performs poorly because it is estimating an interactive effect with a very small time series sample. As PC methods generally

[25] If we did not fix the magnitude of f_T, then the value of non-parallel trending would grow with T.

require $T \to \infty$ and are designed for scenarios with many time observations, we should expect GSC to perform poorly in this situation.

Controlling for a covariate can sometimes alleviate bias for TWFE and GSC. Table 2 shows what happens when treatment is correlated with unobserved factor loadings, including the loadings in the control, but there is no indirect effect of treatment. The covariates help explain some of the non-parallel trending behavior without inducing bias. However, even in this setting, TECCE has a bias that is zero to two decimal points and outperforms both GSC and TWFE in the presence of a random trend. With TECCE, we do not need to take a stance on whether or not treatment affects the outcome through the covariates.

Table 3 presents a best case scenario for TECCE over GSC and TWFE. This specification includes factor loadings that are correlated with treatment status and an indirect effect of $\tau = 1$. Interestingly, when $f_t = 1$, GSC and TWFE without controls performs well in estimating the overall ATT, but have fairly constant biases of -1 when including controls. This fact should be unsurprising when we recall that $y_{i,t}(0)$ is linear in the additive fixed effect and the overall ATT because $\mathbf{x}_{i,t}(0)$ is linear in the additive fixed effect and the indirect ATT. We again demonstrate that including a "bad control" still allows us to estimate the direct ATT. When f_t is set as a random trend, both specifications of GSC and TWFE perform poorly. Neither method adequately captures the indirect transmission of treatment through $\mathbf{x}_{i,t}$. However, unlike these estimators, the TECCE estimator performs well in any of the following scenarios, making it a preferable alternative when one wishes to include covariates but is unsure how the treatment and control distributions change over time.

6 Empirical illustration

Since its membership in the WTO in 2001, China's role in the world economy has grown enormously. As a result, the pro-competitive effects of China's WTO accession have attracted considerable attention, so much so that there is by now a separate strand of literature devoted to them. The bulk of the evidence seems to suggest that both the level and dispersion of markups have gone down following the WTO entry, and that this development has had important welfare effects (Hsu et al., 2020).

The standard approach to studying WTO membership on market characteristics is to exploit differences in tariffs across industries. For example, one can split industries into a treatment and a control group, where the former is relatively more exposed to the WTO accession. Given that pre-WTO tariffs varied greatly across industries, industries that had previously been protected with high tariffs experienced greater tariff reduction. They should therefore be relatively more exposed to the "treatment". The effect of the WTO accession is then estimated via a difference-in-differences-style OLS regression in which markup is regressed onto a dummy variable that takes on the value one for treated industries in post-WTO periods, control variables, and industry and time fixed effects.

While popular, the standard approach to WTO evaluation has (at least) two drawbacks. First, it requires the parallel trends assumption, which may not be realistic in this context. A commonly cited reason is that certain industries have more lobbying power for protection and are thus less responsive to changes in the macroeconomy. Tariffs may be granted to domestic special interest groups, the pressure of which may vary over time (Xiang et al., 2017; Fan et al., 2018; Deng et al., 2018). Differences in lobbying power may therefore cause the treatment and control groups to differ systematically over time, even if China had not joined the WTO in 2001.[26] Because many sources of possible non-parallel trending are unknown and lack good proxies, it is common to control for industry-specific linear time trends (Liu and Qiu, 2016; Mao and Xu, 2019). Deterministic trends can account for some non-parallel trending but not all. Moreover, results tend to be highly sensitive to the inclusion of such trends.[27]

The second main drawback of the standard approach is that it cannot handle covariates that are affected by treatment. This issue is important because the literature has identified many channels through which the WTO accession may affect markups (Brandt et al., 2017; Fan et al., 2018; Deng et al., 2018; Mao and Xu, 2019; Liu and Ma, 2021). Two common examples are the price- and cost-change channels. Markup is defined as the ratio of price to marginal cost. Thus,

[26] Similarly, policymakers may lower tariffs selectively in industries that are able to compete with relatively less expensive imports, for example, in industries experiencing a productivity boom (Brandt et al., 2017).

[27] Some studies include common controls that are thought to be highly correlated with various kinds of protectionism, such as wage rates, employment, exports, and imports (Hsu et al., 2020). Again the results tend to be very sensitive.

markup changes can come from price changes, cost changes, or both. It is therefore common to include one of these variables as a covariate, and to estimate the effect of the WTO accession on it to understand the causal mechanism of treatment (Lu and Yu, 2015; Fan et al., 2018; Mao and Xu, 2019). But then we know from Section 3 that treatment-affected covariates require special treatment or else the estimated ATT will be misleading. Specifically, the inclusion of such covariates will absorb some effect of treatment. The following quotation, taken from Fan et al. (2018, page 116), suggests that researchers are aware of this problem: "If the marginal-cost channel indeed plays a role, then once the marginal costs are included as an explanatory variable, we would witness attenuation of the impact of input tariffs on markups."

The present paper is not the first to point to these shortcomings, but it is the first to consider an econometric approach that is designed to deal with both. The TECCE approach allows for interactive fixed effects in which there may be unobserved differences between cross-sectional units that change over time as a result of common shocks. The parallel trend condition is therefore not required, which is a substantial advantage when compared to the standard fixed effects-based approach. Another advantage of the approach that we exploit in this section is that it not only allows for covariates that may be affected by treatment but that it makes it possible to assess the relative importance of the direct and indirect treatment channels. It should therefore be well suited for the problem at hand.

The data set that we use is taken from Lu and Yu (2015) and comprises 164 industries (three-digit Chinese industrial classification) observed over the 1998–2005 period.[28] The smallness of T here, which is a feature of most data sets in the literature, means that it is important to use techniques that work even if T is not large. The Monte Carlo results reported in Section 5 suggest that the proposed TECCE approach should work well. Following Lu and Yu (2015), the outcome variable is markup dispersion, as measured by the markup Theil index (in logs). Industries are split in half into the treatment and control groups based on whether they faced tariffs above or below the sample median in 2001.

We focus on the Lu and Yu (2015) study in part because of their analysis of the price- and cost-change channels (see their Section E). The authors use the TFP Theil index as a proxy for

[28] See also Deng et al. (2018) who use the same data.

marginal costs. They argue that its inclusion allows them to partially isolate the price-change channel. In order to assess the ATT of the WTO accession on costs, the authors run a second OLS regression with the TFP Theil index as dependent variable and markup dispersion as a covariate. The estimated ATTs are significant, which is taken as evidence that both channels are operational. The purpose of this illustration is to assess whether or not the treatment affects the controls.

The above discussion suggests that in terms of the notation of Section 2, $y_{i,t}$ is the markup Theil index and $\mathbf{x}_{i,t}$ is the TFP Theil index. The estimated factors in $\widehat{\mathbf{f}}_t$ are made up of the cross-sectional averages of these variables. A constant is included as an observed factor (as explained in Section 4), which is tantamount to allowing for industry fixed effects. We therefore allow for one known and two unknown factors. In order to assess if this is enough, we apply the rank classifier of De Vos et al. (2024). The estimated number of factors and rank of the matrix of average factor loadings are equal to one and two, respectively, suggesting that Assumption 7 is met.

Insert Figure 1 about here

The estimated overall and indirect ATTs are reported in Figure 1. The estimates are reported for each year and averaged over all the post- and pre-treatment periods, as is customary in the literature. Both types are reported together with 95% confidence intervals. We first note that both the total and indirect ATT estimates are negative, suggesting that markup dispersion decreased more in industries that had relatively high tariffs in 2001. Given that industries with higher initial tariffs experienced greater tariff reduction, these results imply that the WTO accession reduced markup dispersion. We also note that all pre-treatment estimates are close to zero, which implies that our method adequately captures the non-parallel trending in markups before 2001.

While insignificant in 2002 and 2003, the year-specific total ATTs reported in Figure 1 (a) are significant in 2004 and 2005. The point estimate in 2003 is notably noisy. A possible reason for this is that the industry classification system changed in 2003, as noted by, for example, Lu and Yu (2015) and Chen et al. (2019). The estimated average ATT during the whole post-treatment period is about -0.1 and significant, consistent with the results of Chen et al. (2019).

We also look at the estimated indirect ATT to understand the impact of productivity changes on markups. According to the results reported in Figure 1 (b), the estimated IATTs are negative and significant in the post-treatment period and insignificant in the pre-treatment period. Lu and Yu (2015) estimate the ATT on the TFP Theil index and find it to be significantly negative; however, their approach does not allow them to infer whether this negative response of the TFP Theil index has an effect on the markup Theil index. According to our results, the estimated IATTs are sizable, accounting for almost half of the total ATTs. This result is important in itself, but also for what it means for the results in Lu and Yu (2015), which are based on including the TFP Theil index as a covariate. In particular, we know from before that this type of conditioning will absorb the indirect effect. In this case, since both ATTs are estimated to be negative, and the magnitude of the indirect ATTs are about half of the overall ATTs, conditioning on the TFP Theil index will lead to an underestimation of the total ATTs by about 50%.

7 Conclusion

In this paper, we propose a new estimator of the ATT, dubbed “TECCE”, that is applicable in panels with few time periods when the parallel trends condition fails because of the presence of interactive fixed effects. It can incorporate time- and unit-varying covariates that load on the same factors as the outcome variable. This assumption allows us to use the cross-sectional averages of the observables to estimate the untreated potential outcomes in post-treatment time periods. The covariates are allowed to depend on the treatment status so that TECCE makes it possible separate the direct ATT that is unrelated to the covariates from the indirect ATT that works through those covariates. The estimator is shown to be consistent and asymptotically normal, thereby enabling standard inference, provided only that the number of cross-sectional units, N, is large. This condition is a great advantage in practice because in the literature, many data sets involve only a few time periods T. We consider one such small-T data set in our empirical illustration and estimate the effect of China’s accession into the WTO on the dispersion of industry-level markups. Our results suggest that not adequately capturing changes in productivity leads to drastic underestimation of trade liberalization on markups.

References

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

Aklin, M. and P. Bayer (2017): "How can we estimate the effectiveness of institutions? Solving the post-treatment versus omitted variable bias dilemma," Working Paper.

Angrist, J. D. and J.-S. Pischke (2009): Mostly harmless econometrics: An empiricist's companion, Princeton university press.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021): "Synthetic difference-in-differences," American Economic Review, 111, 4088–4118.

Bai, J. (2009): "Panel data models with interactive fixed effects," Econometrica, 77, 1229–1279.

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

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting event study designs: Robust and efficient estimation," Review of Economic Studies, forthcoming.

Brandt, L., J. Van Biesebroeck, L. Wang, and Y. Zhang (2017): "WTO accession and performance of Chinese manufacturing firms," American Economic Review, 107, 2784–2820.

Breitung, J. and P. Hansen (2021): "Alternative estimation approaches for the factor augmented panel data model with small T," Empirical Economics, 60, 327–351.

Brown, N. and K. Butts (2023): "Dynamic Treatment Effect Estimation with Interactive Fixed Effect Models and Short Panels," Working Paper.

Brown, N. L., P. Schmidt, and J. M. Wooldridge (2022): "Simple alternatives to the common correlated effects model," ArXiv:2112.01486.

Caetano, C. and B. Callaway (2023): "Difference-in-Differences with Time-Varying Covariates in the Parallel Trends Assumption," ArXiv: 2202.02903.

Callaway, B. and S. Karami (2023): “Treatment effects in interactive fixed effects models with a small number of time periods,” Journal of econometrics, 233, 184–208.

Callaway, B. and P. H. Sant’Anna (2021): “Difference-in-Differences with Multiple Time Periods,” Journal of Econometrics, S0304407620303948.

Callaway, B. and E. S. Tsyawo (2023): “Treatment Effects in Staggered Adoption Designs with Non-Parallel Trends,” arXiv preprint arXiv:2308.02899.

Chan, M. K. and S. S. Kwok (2022): “The PCDID approach: difference-in-differences when trends are potentially unparallel and stochastic,” Journal of Business & Economic Statistics, 40, 1216–1233.

Chen, W., X. Chen, C.-T. Hsieh, and Z. Song (2019): “A forensic examination of China’s national accounts,” Brookings Papers on Economic Activity, 77–141.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011): “Weak and strong cross-section dependence and estimation of large panels,” The Econometrics Journal, 14, C45–C90.

de Chaisemartin, C. and X. D’Haultfœuille (2020): “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects,” American Economic Review, 110, 2964–2996.

De Vos, I., G. Everaert, and V. Sarafidis (2024): “A method to evaluate the rank condition for CCE estimators,” Econometric Reviews, 43, 123–155.

Deng, X., R. Jing, and Z. Liang (2018): “Trade liberalisation and domestic brands: Evidence from China’s accession to the WTO,” World Economy, 43, 2237–2262.

Fan, H., X. Gao, Y. A. Li, and T. A. Luong (2018): “Trade liberalization and markups: Micro evidence from China,” Journal of Comparative Economics, 46, 103–130.

Gobillon, L. and T. Magnac (2016): “Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls,” Review of Economics and Statistics, 98, 535–551.

Hayakawa, K. (2016): “Identification problem of GMM estimators for short panel data models with interactive fixed effects,” Economics Letters, 139, 22–26.

Hsu, W.-T., Y. Lu, and G. L. Wu (2020): “Competition, markups, and gains from trade: A quantitative analysis of China between 1995 and 2004,” Journal of International Economics, 122, 103266.

Huber, M. (2014): “Identifying causal mechanisms (primarily) based on inverse probability weighting,” Journal of Applied Econometrics, 29, 920–943.

Juodis, A., H. Karabiyik, and J. Westerlund (2021): “On the Robustness of the Pooled CCE Estimator,” Journal of Econometrics, 220, 325–348.

Kim, D. and T. Oka (2014): “Divorce Law Reforms and Divorce Rates in the USA: An Interactive Fixed-Effects Approach,” Journal of Applied Econometrics, 29, 231–245.

Leeb, H. and B. M. Pötscher (2005): “Model selection and inference: Facts and fiction,” Econometric Theory, 21, 21–59.

Liu, Q. and L. D. Qiu (2016): “Intermediate input imports and innovations: Evidence from Chinese firms’ patent filings,” Journal of International Economics, 103, 166–6–183.

Liu, Z. and H. Ma (2021): “Input trade liberalisation and markup distribution: Evidence from China,” Economic Inquiry, 59, 344–360.

Lu, Y. and L. Yu (2015): “Trade liberalization and markup dispersion: evidence from China’s WTO accession,” American Economic Journal: Applied Economics, 7, 221–253.

Mao, Q. and J. Xu (2019): “Input trade liberalisation, institution and markup: Evidence from China’s accession to the WTO,” World Economy, 42, 3537–3568.

Moon, H. R. and M. Weidner (2015): “Linear regression for panel with unknown number of factors as interactive fixed effects,” Econometrica, 83, 1543–1579.

——— (2019): “Nuclear Norm Regularized Estimation of Panel Regression Models,” Working Paper.

Pesaran, M. H. (2006): “Estimation and inference in large heterogeneous panels with a multifactor error structure,” Econometrica, 74, 967–1012.

Sant'Anna, P. H. and J. Zhao (2020): "Doubly Robust Difference-in-Differences Estimators," Journal of Econometrics, 219, 101–122.

Westerlund, J., Y. Petrova, and M. Norkute (2019): "CCE in fixed-T panels," Journal of Applied Econometrics, 34, 746–761.

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

Xiang, X., F. Chen, C.-Y. Ho, and W. Yue (2017): "Heterogeneous effects of trade liberalisation on firm-level markups: Evidence from China," World Economy, 40, 1667–1686.

Xu, Y. (2017): "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," Political Analysis, 25, 57–76.

Table 1: Monte Carlo results with $\kappa = 0$ and $\tau = 0$.

			GSC		TWFE	
N	T	TECCE	No Controls	$+ \mathbf{x}_{i,t}$	No Controls	$+ \mathbf{x}_{i,t}$
		Factor specification: $f_t = 1$				
50	5	-0.00	-0.00	-0.00	-0.00	-0.00
		[0.241]	[0.178]	[0.127]	[0.181]	[0.129]
50	15	0.01	0.01	0.00	0.00	-0.00
		[0.166]	[0.156]	[0.114]	[0.159]	[0.118]
300	5	-0.00	-0.00	-0.00	-0.00	-0.00
		[0.094]	[0.074]	[0.052]	[0.076]	[0.053]
300	15	0.00	0.00	0.00	0.00	0.00
		[0.072]	[0.067]	[0.046]	[0.069]	[0.047]
		Factor specification: $f_t = 1 + t/8$				
50	5	-0.00	-0.01	0.00	-0.01	-0.00
		[0.235]	[0.186]	[0.130]	[0.212]	[0.140]
50	15	-0.01	-0.00	0.00	-0.02	0.00
		[0.183]	[0.169]	[0.116]	[0.367]	[0.206]
300	5	-0.00	0.00	-0.00	0.00	0.00
		[0.104]	[0.072]	[0.052]	[0.084]	[0.057]
300	15	-0.00	-0.00	0.00	0.00	0.00
		[0.072]	[0.068]	[0.050]	[0.152]	[0.086]
		Factor specification: $f_t = 1.5 + v_t$				
50	5	-0.01	-0.00	0.01	-0.00	0.01
		[0.247]	[0.177]	[0.124]	[0.186]	[0.134]
50	15	0.00	-0.00	0.00	-0.01	0.00
		[0.163]	[0.153]	[0.108]	[0.537]	[0.287]
300	5	-0.01	-0.00	0.00	0.01	0.01
		[0.125]	[0.085]	[0.058]	[0.205]	[0.110]
300	15	0.00	0.00	0.00	0.00	-0.00
		[0.080]	[0.075]	[0.052]	[0.306]	[0.158]

Notes: This table reports the bias and RMSE under different simulation specifications. Each row presents the bias and the RMSE in brackets below. The TECCE estimator is this paper's proposed estimator. GSC corresponds to the generalized synthetic control estimator from Xu (2017). TWFE corresponds to the imputation estimator of Borusyak et al. (2024) which uses a two-way fixed effect model. The column labeled No Controls does not control linearly for x_{it} while the column labeled $+x_{it}$ does. 1,500 simulation draws.

Table 2: Monte Carlo results with $\kappa = -0.5$ and $\tau = 0$.

			GSC		TWFE	
N	T	TECCE	No Controls	+ $\mathbf{x}_{i,t}$	No Controls	+ $\mathbf{x}_{i,t}$
Factor specification: $f_t = 1$						
50	5	0.00	0.01	0.00	0.01	0.00
		[0.220]	[0.157]	[0.111]	[0.181]	[0.127]
50	15	0.01	0.01	0.01	0.01	0.01
		[0.147]	[0.144]	[0.100]	[0.160]	[0.114]
300	5	0.01	0.00	-0.00	0.00	-0.00
		[0.144]	[0.065]	[0.044]	[0.074]	[0.050]
300	15	0.00	0.00	-0.00	-0.00	-0.00
		[0.062]	[0.059]	[0.043]	[0.068]	[0.048]
Factor specification: $f_t = 1 + t/8$						
50	5	0.01	0.01	0.00	-0.30	-0.14
		[0.252]	[0.163]	[0.119]	[0.366]	[0.202]
50	15	0.00	0.00	0.01	-0.94	-0.33
		[0.156]	[0.146]	[0.102]	[1.011]	[0.387]
300	5	0.00	0.00	0.00	-0.31	-0.14
		[0.095]	[0.064]	[0.046]	[0.323]	[0.155]
300	15	-0.00	0.00	0.00	-0.94	-0.34
		[0.062]	[0.059]	[0.041]	[0.952]	[0.348]
Factor specification: $f_t = 1.5 + v_t$						
50	5	-0.00	-0.00	0.00	0.67	0.22
		[0.201]	[0.145]	[0.103]	[0.733]	[0.273]
50	15	-0.01	-0.01	-0.00	0.86	0.28
		[0.152]	[0.141]	[0.099]	[0.923]	[0.339]
300	5	0.00	-0.00	-0.00	0.39	0.14
		[0.106]	[0.058]	[0.042]	[0.404]	[0.150]
300	15	-0.00	-0.00	0.00	1.42	0.37
		[0.061]	[0.057]	[0.039]	[1.440]	[0.389]

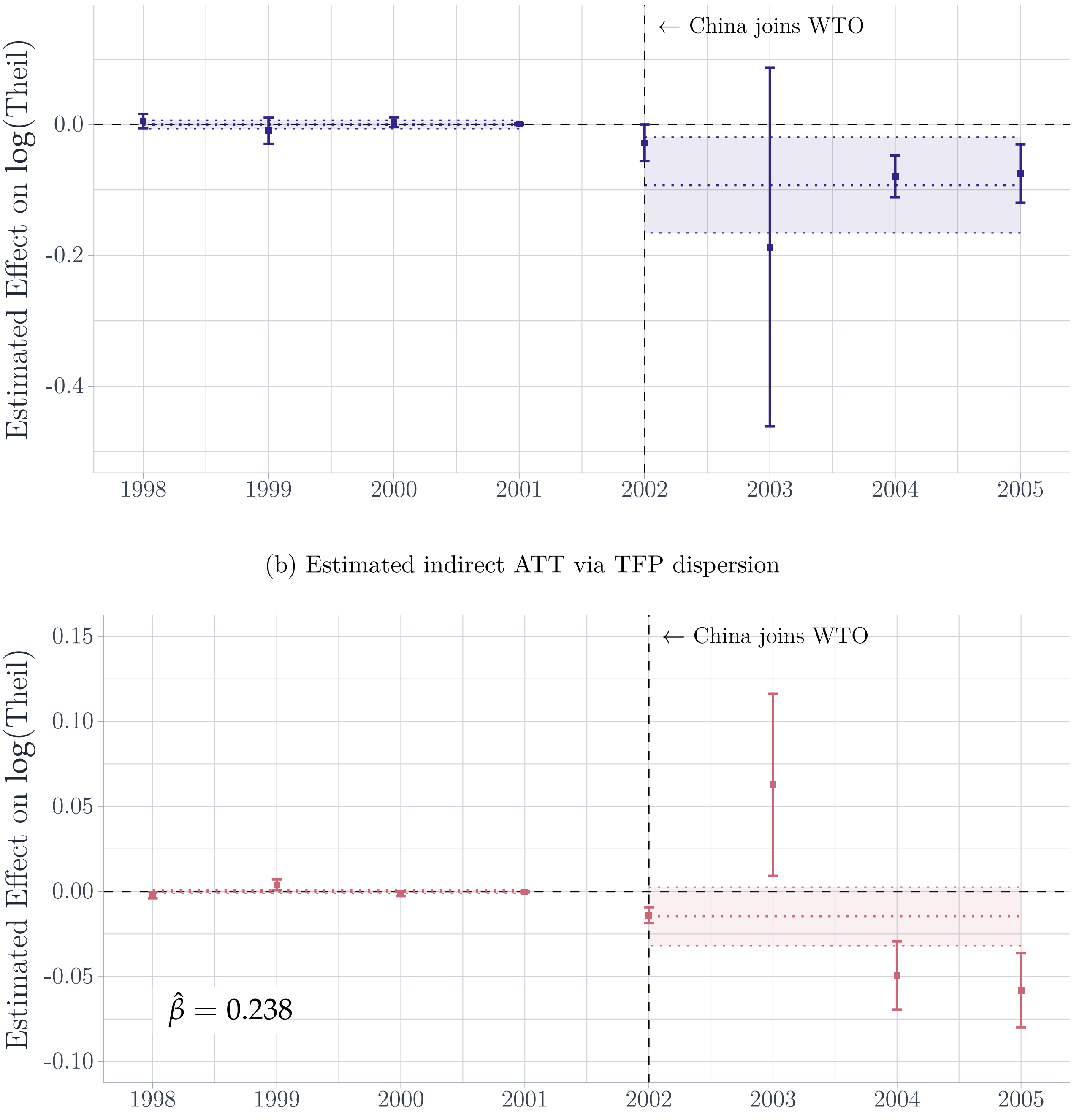
Notes: This table reports the bias and RMSE under different simulation specifications. Each row presents the bias and the RMSE in brackets below. The TECCE estimator is this paper's proposed estimator. GSC corresponds to the generalized synthetic control estimator from Xu (2017). TWFE corresponds to the imputation estimator of Borusyak et al. (2024) which uses a two-way fixed effect model. The column labeled No Controls does not control linearly for x_{it} while the column labeled $+x_{it}$ does. 1,500 simulation draws.

Table 3: Monte Carlo results with $\kappa = -0.5$ and $\tau = 1$.

			GSC		TWFE	
N	T	TECCE	No Controls	$+ \mathbf{x}_{i,t}$	No Controls	$+ \mathbf{x}_{i,t}$
		Factor specification: $f_t = 1$				
50	5	0.01	-0.01	-1.05	-0.01	-1.01
		[0.206]	[0.152]	[1.057]	[0.175]	[1.017]
50	15	0.00	0.00	-1.03	0.00	-1.00
		[0.150]	[0.144]	[1.036]	[0.165]	[1.007]
300	5	0.00	-0.00	-1.03	-0.00	-1.00
		[0.092]	[0.063]	[1.036]	[0.072]	[1.000]
300	15	0.00	0.00	-1.03	0.00	-1.00
		[0.064]	[0.060]	[1.029]	[0.068]	[1.000]
		Factor specification: $f_t = 1 + t/8$				
50	5	0.01	0.01	-1.26	-0.30	-1.20
		[0.206]	[0.158]	[1.273]	[0.369]	[1.209]
50	15	-0.00	-0.00	-1.13	-0.95	-1.62
		[0.159]	[0.147]	[1.139]	[1.022]	[1.640]
300	5	0.00	0.00	-1.25	-0.31	-1.20
		[0.093]	[0.066]	[1.251]	[0.323]	[1.205]
300	15	0.00	-0.00	-1.13	-0.94	-1.62
		[0.066]	[0.062]	[1.131]	[0.952]	[1.618]
		Factor specification: $f_t = 1.5 + v_t$				
50	5	0.00	0.00	-1.33	0.08	-1.31
		[0.214]	[0.161]	[1.339]	[0.203]	[1.320]
50	15	0.00	0.00	-1.09	-2.17	-2.03
		[0.175]	[0.165]	[1.094]	[2.301]	[2.069]
300	5	0.00	-0.00	-1.19	-0.14	-1.24
		[0.096]	[0.066]	[1.194]	[0.157]	[1.243]
300	15	-0.00	-0.00	-1.11	-1.20	-1.76
		[0.064]	[0.060]	[1.112]	[1.216]	[1.767]

Notes: This table reports the bias and RMSE under different simulation specifications. Each row presents the bias and the RMSE in brackets below. The TECCE estimator is this paper's proposed estimator. GSC corresponds to the generalized synthetic control estimator from Xu (2017). TWFE corresponds to the imputation estimator of Borusyak et al. (2024) which uses a two-way fixed effect model. The column labeled No Controls does not control linearly for x_{it} while the column labeled $+x_{it}$ does. 1,500 simulation draws.

Figure 1: Estimated ATTs of China's WTO accession in 2001 on the markup Theil index.



(a) Estimated total ATT

(b) Estimated indirect ATT via TFP dispersion

Notes: The figures present ATT estimates and 95% confidence intervals for the effect of China's WTO accession in 2001 on the dispersion of markups as measured by the markup Theil index. The treatment group comprise all industries that in 2001 had above-median tariff rates. Estimates are computed using the TECCE estimator with the TFP Theil index as a covariate. A constant is included as an observed factor. Figure (a) presents estimates of the total ATT and figure (b) presents the estimated indirect ATT operating through the TFP Theil index. $\hat{\beta}$ in figure (b) refers to the estimated slope on the TFP Theil index in the markup Theil index regression.