DEUTSCHE
BUNDESBANK
EUROSYSTEM

# Discussion Paper

Deutsche Bundesbank
No 30/2023

## Staggered difference-in-differences in gravity settings: Revisiting the effects of trade agreements

Arne J. Nagengast
(Deutsche Bundesbank)

Yoto V. Yotov
(Drexel University, ifo and CESifo)

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de



ISBN 978–3–95729–965–9
ISSN 2941–7503

# Non-technical summary

### Research Question

The effect of regional trade agreements (RTAs) on trade flows is among the most widely studied and debated topics in the international trade literature. Recent analyses of the impact of RTAs obtain estimates that are perceived as unrealistically small. In addition, a series of papers provide evidence for significant heterogeneity in the effects of RTAs across agreements. However, a recent econometric literature has demonstrated that the conventionally-used estimators can be severely biased in this setting and has proposes estimators that are heterogeneity-robust.

### Contribution

Against this backdrop, the contribution our paper is to adapt and nest the new heterogeneity-robust methods within a so-called gravity model, which is commonly used to analyze bilateral flows such as trade, foreign direct investment, migration, etc. The empirical gravity model that we use also takes into account the best estimation practices from the trade gravity literature. Based on our analysis of the impact of RTAs, we expect the new estimation approach to have significant implications for the estimates of other policies that have been studied with the gravity model, e.g., economic sanctions, membership in GATT and WTO, and currency unions.

### Results

The main result from our analysis is that the new estimation approach delivers RTA estimates that are significantly larger and longer than corresponding estimates that are based on the current methods from the gravity literature. Importantly, the larger estimates of the effects of RTAs are also more plausible from a policy perspective. The conventional estimator is biased downward since it puts larger weights on short-run effects and on those of late cohorts, which are both associated with smaller treatment effects.

# Nichttechnische Zusammenfassung

### Fragestellung

Die Auswirkungen regionaler Handelsabkommen (RTAs) auf internationale Handelsströme gehören zu den am meisten untersuchten und kontrovers diskutierten Themen in der internationalen Handelsliteratur. Die Schätzwerte jüngster Analysen der Auswirkungen von RTAs werden als unrealistisch klein angesehen. Darüber hinaus gibt eine Reihe von Arbeiten Hinweise auf eine erhebliche Heterogenität der Auswirkungen von RTAs über verschiedene Abkommen hinweg. Eine neuere ökonometrische Literatur hat jedoch gezeigt, dass die herkömmlich verwendeten Schätzer vor diesem Hintergrund stark verzerrt sein können und hat heterogenitätsrobuste Alternativen vorgeschlagen.

### Beitrag

Vor diesem Hintergrund besteht der Beitrag unseres Papiers darin, die neuen heterogenitätsrobusten Methoden an ein sogenanntes Gravitationsmodell anzupassen und darin zu integrieren, welches üblicherweise zur Analyse bilateraler Ströme wie Handel, ausländische Direktinvestitionen, Migration usw. verwendet wird. Das von uns verwendete empirische Gravitationsmodell berücksichtigt zudem die besten Schätzverfahren aus der internationalen Handelsliteratur. Auf der Grundlage unserer Analyse der Auswirkungen von Freihandelsabkommen gehen wir davon aus, dass der neue Schätzansatz erhebliche Auswirkungen auf die Schätzungen anderer politischer Maßnahmen haben wird, die mit dem Gravitationsmodell untersucht wurden, z. B. Wirtschaftssanktionen, Mitgliedschaft im GATT und in der WTO sowie Währungsunionen.

### Ergebnisse

Das wichtigste Ergebnis unserer Analyse ist, dass der neue Schätzsansatz in deutlich größeren und länger anhaltenden Schätzwerten für RTAs resultiert als entsprechende Schätzungen, die auf den aktuellen Methoden der Gravitationsliteratur beruhen. Hervorzuheben ist, dass die größeren Schätzwerte der Auswirkungen von RTAs auch aus einer politischen Perspektive plausibler sind. Der konventionelle Schätzer ist nach unten verzerrt, da er die kurzfristigen Effekte und die der späten Kohorten stärker gewichtet, die beide mit geringeren Behandlungseffekten verbunden sind.

Deutsche Bundesbank Discussion Paper No 30/2023

# Staggered Difference-in-Differences in Gravity Settings: Revisiting the Effects of Trade Agreements*

Arne J. Nagengast
Deutsche Bundesbank

Yoto V. Yotov
Drexel University
ifo, CESifo

**Abstract**

We nest an extended two-way fixed effect (ETWFE) estimator for staggered difference-in-differences within the structural gravity model. To test the ETWFE, we estimate the effects of regional trade agreements (RTAs). The results suggest that RTA estimates in the current gravity literature may be biased downward (by more than 50% in our sample). Sensitivity analyses confirm the robustness of our main findings and demonstrate the applicability of our methods in different settings. We expect the ETWFE methods to have significant implications for the estimates of other policy variables in the trade literature and for gravity regressions on migration and FDI flows.

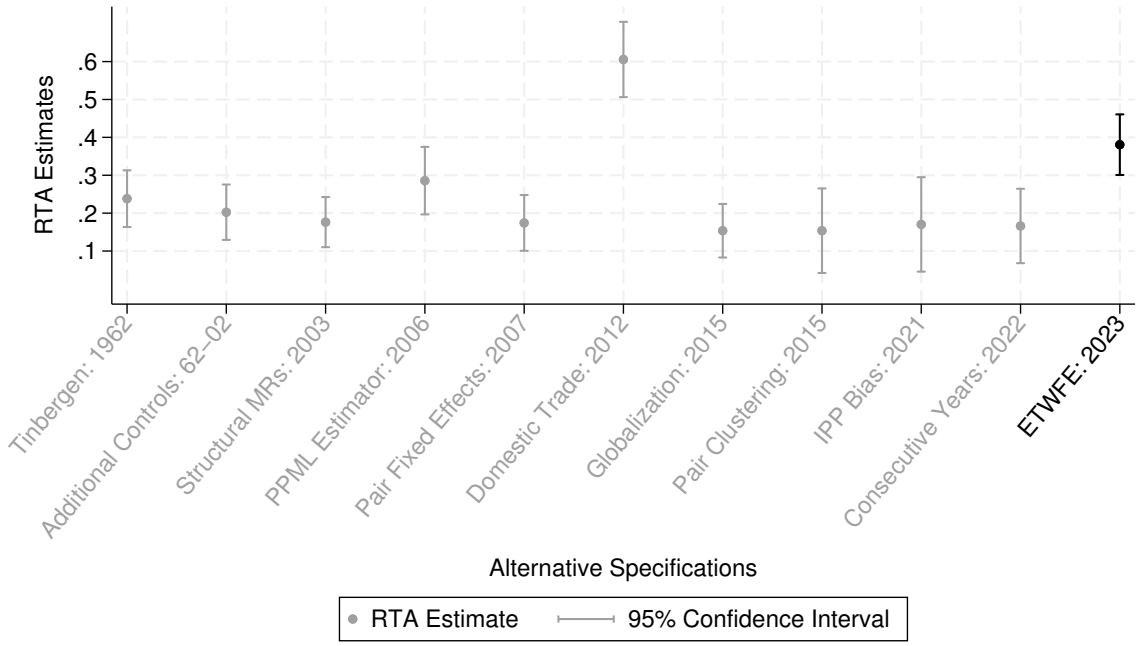
**Keywords:** Staggered Difference-in-Differences, Gravity Model, Trade Agreements
**JEL classification:** C13, C23, F10, F13, F14.

*We would like to thank Xavier Jaravel, Chris Walkers, Jeffrey Wooldridge, and two anonymous referees for insightful comments that greatly improved the paper. We are also grateful to Sebastien Bradley, Peter Egger, Teresa Harrison, Ohyun Kwon, Christopher Meissner, Ninon Moreau-Kastler, Fernando Rios-Avila, Farid Toubal, Thomas Zylkin, and seminar participants at the Deutsche Bundesbank and the Economics of Free Trade Agreements conference at the University of Surrey for their helpful comments and suggestions. All remaining errors are our own. The views expressed in this paper are those of the author(s) and do not necessarily coincide with the views of the Deutsche Bundesbank or the Eurosystem. Contact information: Nagengast – arne.nagengast@bundesbank.de. Yotov – yotov@drexel.edu.

# 1 Introduction

We show that accounting for heterogeneous effects of regional trade agreements (RTAs) across space and time leads to estimates of RTA effects on trade that are twice as large as previously thought. To highlight our contribution, Figure 1 capitalizes on the analysis from Larch and Yotov (2023), who offer a chronological review of the developments in the RTA literature over the past 60 years; from the first, atheoretical gravity specification of

Figure 1: 61 Years of RTA Estimates with the Gravity Model of Trade




**Notes**: This figure replicates some of the specifications from Figure 1 of Larch and Yotov (2023) with our data. The figure plots estimates of the effects of RTAs, along with the corresponding 95% confidence intervals. The X-axis of the figure lists the alternative specifications from the existing literature, which follow the evolution of the estimation methods from the 1962 specification of 'Tinbergen' to the most recent developments that are taken into account in specification 'Consecutive Years'. The last estimate in the figure is obtained with the extended two-way fixed effect (ETWFE) estimator that we implement in this paper. See Section 4 for further details.

Tinbergen (1962), which is labeled 'Tinbergen: 1962', until the most recent developments in the empirical trade literature, as captured by specification 'Consecutive Years: 2022', which comes from Egger et al. (2022).[1]

[1]To construct Figure 1, we use the bilateral trade dataset of the World Trade Organization (Larch et al., 2019a), downloadable at https://www.wto.org/english/res_e/reser_e/structural_gravity_e.htm, and we implement some of the specifications from Larch and Yotov (2023). We offer further details on the sources and construction of our data in Section 3. Specification 'Consecutive Years: 2022' incorporates the current best practice recommendations from the empirical trade literature, including the Poisson

We draw three conclusions from Figure 1. First, from a methodological perspective, the main message is that there have been many and impactful developments in the related literature. Second, from a policy perspective, our estimates suggest that the effects of RTAs on trade have been positive and statistically significant, but relatively small. The most current estimate ('Consecutive Years: 2022') implies that, all else equal, the RTAs in our sample have led to an increase of trade among the RTA members of about 18.1%.[2] Lastly, based on the evolution of the estimates in Figure 1, it seems that the existing literature has converged toward a stable RTA estimate of about 0.17 – 0.18.

Our main result appears last in Figure 1. The specification that we use to obtain it is identical to the preferred trade model ('Consecutive Years: 2022') but, in addition, it implements the latest heterogeneity-robust innovations in the difference-in-differences (DiD) literature on staggered treatment adoption, i.e., when the treatment or intervention occurs in multiple units in different time periods (e.g., Borusyak and Jaravel, 2017; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; Borusyak et al., 2023; de Chaisemartin and D'Haultfoeuille, 2022; Wooldridge, 2023). The new RTA estimate is still positive and statistically significant, however, it is significantly larger (more than double), implying that the RTAs in our sample have led to an increase of about 46% in trade among member countries. Comparisons among the alternative estimates in Figure 1 reveal that, together with the use of domestic trade, the impact of the new estimator on the RTA estimate is among the most impactful methods in relative terms.[3]

From a methodological perspective, our contribution is to adapt and 'nest' the new (heterogeneity-robust) staggered difference-in-differences approach within an empirical

pseudo maximum likelihood (PPML) estimator, exporter-time and importer-time fixed effects, pair fixed effects, domestic trade flows, globalization effects, consecutive-year data, and pair-clustered standard errors. We briefly review the current methods of estimating the gravity model in Section 2.1.

[2]Calculated as $(\exp(0.166) - 1) \times 100 = 18.057$. This estimate is comparable to recent results from the literature. For example, Baier et al. (2019) obtain an estimate of the impact of free trade agreements of 0.19, implying a corresponding trade volume effect of about 21%. The preferred estimate from the survey of Larch and Yotov (2023) is similar (0.18).

[3]Another notable result from our analysis is based on an event study, which we perform in Section 4.3, and it reveals that the new RTA estimates are not only larger but also spread over a longer period of time. A battery of sensitivity experiments, which are based on insights from the recent DiD literature and from the empirical trade literature, confirm the robustness of our main findings.

panel gravity model,[4] which takes into account the best estimation practices from the trade gravity literature, including the use of the non-linear PPML estimator (Santos Silva and Tenreyro, 2006) and multi-dimensional fixed effects (Yotov et al., 2016). While there is an array of excellent recent heterogeneity-robust DiD estimators for linear settings,[5] treatments of non-linear DiD with staggered interventions, e.g., such as PPML, are limited. Therefore, to implement our methods, we rely on Wooldridge (2023), who generalizes the linear setting from Wooldridge (2021), which proposes an *extended two-way fixed effect (ETWFE) estimator*, for the estimation of non-linear models.[6] The ETWFE estimator maintains the general structure of the basic two-way fixed effect (TWFE) regression, but it also allows for treatment effect heterogeneity by additionally introducing suitable cohort (i.e., all units treated in a particular year) and year interactions.

In terms of the existing literature, our contribution is most closely related to the gravity RTA literature, where the 'average treatment effect' methods of Baier and Bergstrand (2007) have become the standard tool to estimate the effects of RTAs. In addition, and more important for our purposes, recent trade papers have shown that the average estimates of the effects of trade agreements mask significant heterogeneity both across treatment groups and also across different time periods, e.g., Baier et al. (2019) and Larch and Yotov (2023). At the same time, estimation methods similar to the ones commonly used in the gravity setting have recently come under considerable scrutiny from econometric papers that analyze settings with staggered DiD designs, where the estimates are possibly biased in the presence of treatment effect heterogeneity, e.g., due to so-called 'forbidden comparisons' that (mis)use already-treated units in the control group (e.g., Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020).

[4]The gravity equation, e.g., Eaton and Kortum (2002) and Anderson and van Wincoop (2003), is the workhorse model of international trade, and it has been used in tens of thousands of applications.

[5]See, for example, Callaway and Sant'Anna (2021), Sun and Abraham (2021), Wooldridge (2021), Borusyak et al. (2023), and de Chaisemartin and D'Haultfoeuille (2022). de Chaisemartin and D'Haultfœuille (2022) and Roth et al. (2023) offer surveys of this literature.

[6]Strictly speaking it is the estimation *model* rather than the estimation *method* that is extended, i.e., the estimation method remains the same, but is applied to a more flexible model. We offer a more detailed discussion of the implementation and implications of ETWFE in Section 2. We thank Jeff Wooldridge for suggesting this clarification.

We also see two broader implications of our staggered difference-in-differences methods of estimating non-linear gravity models with high-dimensional fixed effects. First, in relation to the existing trade literature, the proposed methods can be used to revisit and improve the estimates from a series of policy applications that have been of significant interest to trade economists, including the effects of membership in the World Trade Organization (WTO), e.g., Rose (2004) and Larch et al. (2019a), currency unions, e.g., Rose (2000) and Larch et al. (2019b), and economic sanctions, e.g., Hufbauer and Oegg (2003) and Felbermayr et al. (2020). Second, we believe that our methods can improve on gravity specifications of bilateral flows beyond trade, including migration, e.g., Anderson (2011), foreign direct investment, e.g., Anderson et al. (2019), commuting patterns, e.g., Persyn and Torfs (2016), and, more generally, to any structural or atheoretical setting of bilateral flows that is subject to economic gravity forces.

The rest of the paper is structured as follows. Section 2 describes our methods. Section 3 provides information on our data and the data sources that we use. Section 4 presents our main findings. Section 5 describes the results from a series of robustness analyses and sensitivity experiments that we performed. Section 6 concludes.

# 2 Methods

The objective of this section is to 'nest' the insights from the new heterogeneity-robust staggered difference-in-differences methods within a structural gravity model. To this end, in Subsection 2.1, we review the current best practices to estimate the gravity model of trade. Then, departing from the benchmark gravity setting, in Subsection 2.2, we implement and discuss the implications of the relevant developments in the recent staggered DiD literature. For expositional purposes, and consistent with our empirical analysis, our focus in this section will be on the effects of RTAs.

## 2.1 Estimating structural gravity: A brief review

Our departing point is the following econometric panel gravity model, which takes into account the current best-practice recommendations from the existing literature:[7]

$$Y_{ij,t} = \exp\left[\delta_{ij,t}^{\text{TWFE}} RTA_{ij,t} + \pi_{i,t} + \chi_{j,t} + \tau_{ij} + \theta_{ii,t}\right] \times \epsilon_{ij,t}. \tag{1}$$

Here, $Y_{ij,t}$ denotes nominal trade flows (Baldwin and Taglioni, 2006) from country $i$ to country $j$ at time $t$, including domestic trade flows (Yotov, 2022). To account for heteroskedasticity and to take advantage of the information contained in zero trade flows, we use the PPML estimator (Santos Silva and Tenreyro, 2006, 2021). $\pi_{i,t}$ are exporter-year fixed effects and $\chi_{j,t}$ are importer-year fixed effects, which account for the multilateral resistance terms of Anderson and van Wincoop (2003) as well as for any other country-time determinants of trade flows on the exporter and on the importer side (Hummels, 2001; Baldwin and Taglioni, 2006; Olivero and Yotov, 2012). $\tau_{ij}$ are directional pair fixed effects (Baier et al., 2019), which mitigate endogeneity concerns (Baier and Bergstrand, 2007), and would control for all symmetric and asymmetric time-invariant trade costs (Egger and Nigai, 2016; Agnosteva et al., 2019). $\theta_{ii,t}$ denotes a set of dummy variables that are equal to one for domestic trade and zero for international trade for each year in the sample.[8] These covariates would account for common globalization trends, e.g., improvements in communication, transportation, etc., (Bergstrand et al., 2015). Lastly, $\epsilon_{ij,t}$ is a multiplicative error term, and we will cluster all standard errors by country-pair (Egger and Tarlea, 2015).[9]

Most important for our purposes, $\delta_{ij,t}^{\text{TWFE}}$ is the TWFE estimate of the effect of RTAs on trade, whose estimate is of central interest to us. While, from a structural gravity

[7] For brevity, we only offer motivation and representative references for each of the terms in our gravity specification (1). We refer the reader to Larch and Yotov (2023) for a complete bibliography and a detailed discussion of the evolution of the gravity methods with a focus on RTAs.

[8] For example, $\theta_{ii,2000}$ is a dummy variables that takes the value of one for domestic trade in 2000 and otherwise zero.

[9] In some robustness checks, we also additionally include a vector of gravity controls such as distance, contiguity, language, colonial ties, WTO membership, GDP, and remoteness indexes. We refer the reader to Yotov et al. (2016) for a discussion of the alternative practices for gravity estimations.

perspective, the estimate of the effects of RTAs could vary across the same dimensions as the bilateral trade flows data itself, i.e., '$i, j, t$',[10] the vast majority of RTA applications impose a common/average effect of RTAs (Baier and Bergstrand, 2007; Anderson and Yotov, 2016). Only recently, have there been some attempts to allow for certain heterogeneity in the effects of RTAs, e.g., across agreements, pairs, or time (Baier et al., 2019; Larch and Yotov, 2023). However, as noted earlier, the existing RTA estimates from the gravity literature are possibly biased in the presence of treatment effect heterogeneity, e.g., due to so-called 'forbidden comparisons' that (mis)use already-treated units in the control group (Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020), and this is the main motivation for our current contribution and analysis.

## 2.2 Gravity with heterogeneity-robust staggered DiD

The TWFE estimator has recently come under considerable scrutiny in settings with staggered DiD designs, i.e., in which an intervention occurs in multiple units in different time periods. In the presence of heterogenous treatment effects, i.e., if the effect of the intervention is heterogeneous between groups or over time, which is exactly the case with many of the policies that are evaluated with the gravity model, the TWFE estimator may result in biased estimates that are difficult to interpret. As shown by Borusyak and Jaravel (2017) and de Chaisemartin and D'Haultfœuille (2020), TWFE estimates are a weighted sum of average treatment effects on the treated (ATTs) in each group and time period, with weights that may be negative. As a consequence, TWFE regressions may not identify a convex combination of treatment effects, i.e., the TWFE coefficient may be negative even though all individual ATTs are positive. The source of the bias lies in so-called 'forbidden comparisons', in which units that are treated in early periods

[10]The reason is that, regardless of whether the gravity model is derived on the demand side, e.g., Anderson and van Wincoop (2003) or on the supply side, e.g., Eaton and Kortum (2002), the coefficient on the RTA variable is a function of two parameters – the direct elasticity of trade with respect to RTAs and the trade elasticity. As demonstrated by Carrere et al. (2020), the latter can vary across the dimensions of the trade data. Thus, the composite elasticity $\delta_{ij,t}^{\text{TWFE}}$ should vary across the same dimensions. We thank an anonymous referee for asking us to add a discussion about the dimensionality of the RTA estimate from a structural gravity perspective.

(already-treated units) are included in the control group for units treated in later periods (Borusyak and Jaravel, 2017).

For the linear setting, a wide range of heterogeneity-robust DiD estimators have been proposed in the recent literature (e.g., de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; Borusyak et al., 2023), which differ inter alia in their identifying assumptions, comparison groups, and efficiency properties.[11] An explicit treatment of non-linear difference-in-differences with staggered interventions, like the gravity setting considered in equation (1), is more limited. For the estimation of non-linear models, Wooldridge (2023) proposes an extended TWFE estimator analogous to the linear case considered in Wooldridge (2021). The extended TWFE estimator maintains the general structure of the basic TWFE regression, but allows for treatment effect heterogeneity at the cohort-year level by additionally introducing suitable cohort and calendar year interactions.[12]

We preserve all recommendations from the gravity literature and, following Wooldridge (2023), we make one important adjustment to equation (1). Specifically, we replace the single indicator variable for the presence of an RTA between $i$ and $j$ at time $t$ ($RTA_{ij,t}$) with the following term:

$$\sum_{g=q}^{T}\sum_{s=g}^{T}\delta_{gs}D_{gs}, \tag{2}$$

where country pair $ij$ belongs to treatment cohort $g$ if the RTA onset was in year $g$, $q$ is the first year of the treatment of cohort $g$, $T$ is the last year of the panel, $D_{gs}$ is a time-

[11] See, for example, de Chaisemartin and D'Haultfœuille (2022) and Roth et al. (2023) for recent surveys of the literature. See also, for example, Cengiz et al. (2019), Deshpande and Li (2019), Egger et al. (2021), and Gardner (2022) for stacked regression approaches.

[12] Note that, like many other estimators in the literature (e.g., de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021), the extended TWFE with cohort-year interactions, in principle, allows for arbitrary heterogeneity within cohort-year cells, but can only identify (simple) average treatment effects at the cohort-year level. If one is interested in different (weighted) averages of cohort-year effects or even in individual treatment effects, additional interactions or an imputation approach are necessary (Borusyak et al., 2023). We offer further related discussion in Section 5, where we also relax these restrictions by considering an ETWFE specification, which allows for heterogeneity at the agreement-year level, as well as a Poisson imputation estimator (Wooldridge, 2023).

varying treatment indicator equal to 1 for cohort $g$ for $s = t$ in post-treatment years and 0 otherwise, and $\delta_{gs}$ captures the cohort-year specific treatment effects.[13] Equation (2) avoids comparisons with already-treated units by saturating the respective observations with fixed effects.[14]

Taking into account the above considerations, our main estimating equation becomes:

$$Y_{ij,t} = \exp\left[\sum_{g=q}^{T}\sum_{s=g}^{T}\delta_{gs}D_{gs} + \pi_{i,t} + \chi_{j,t} + \tau_{i,j} + \theta_{ii,t}\right] \times \epsilon_{ij,t}. \tag{3}$$

As noted earlier, the only difference between equations (3) and (1) is our treatment of RTAs. In the empirical analysis, we will obtain and compare the corresponding estimates from the TWFE and ETWFE specifications.[15] Before that, we explicitly state under which conditions the treatment effects of interest are identified (Section 2.2.1), discuss the aggregation of the cohort-year specific treatment effects (Section 2.2.2), and illustrate the potential bias of the TWFE with an example (Section 2.2.3).

### 2.2.1 Identifying assumption

Initially, we are interested in identifying cohort-time-specific (approximate) proportional treatment effects. For brevity, in the following, we will often refer to these simply as 'treatment effects'. Inference on ATTs, i.e., the level effects, is somewhat complicated in this setting by the need to obtain consistent coefficient estimates including standard errors for all fixed effects in equation (3) (Wooldridge, 2023). Following Wooldridge (2023) in the potential outcome notation of Athey and Imbens (2022), the proportional treatment

[13] For example, $\delta_{1990,1990}$ captures the treatment effect of all RTAs signed in the year 1990 on trade in the first year, $\delta_{1990,1991}$ captures the treatment effect of the same RTAs on trade in the second year, etc.

[14] Note that in the current form, equation (2) includes both not-yet treated and never treated units as controls (e.g., Callaway and Sant'Anna, 2021), though the comparison group can also be restricted (see also Section 5).

[15] In terms of practical implementation, we augmented the Stata command *jwdid* (Rios-Avila, 2022) to account for the best estimation practices from the gravity literature, which we summarized in Section 2.1.

effects are for all $g \in \{q, \ldots, T\}$ and $s \in \{g, \ldots, T\}$

$$\delta_{gs} = \ln\Big(\frac{(E[Y_s(g)|D_g = 1])}{(E[Y_s(\infty)|D_g = 1])}\Big) = \ln(E[Y_s(g)|D_g = 1]) - \ln(E[Y_s(\infty)|D_g = 1]), \quad (4)$$

where $Y_t(g)$ is the potential outcome of cohort $g$ at time $t$, $Y_t(\infty)$ is the potential outcome in the never treated state, i.e., of units not subjected to the treatment over a particular time period, and $D_g$ is a binary variable indicating membership to cohort $g$. To identify $\delta_{gs}$, two assumptions are necessary, which we explicitly test in Section 4.1.[16]

First, the no anticipation assumption requires for all $g \in \{q, \ldots, T\}$ and $t \in \{1, \ldots, g-1\}$ that

$$E[Y_t(g)|D_g = 1] = E[Y_t(\infty)|D_g = 1, \mathbf{X}], \quad (5)$$

where $\mathbf{X}$ is a vector of covariates subsuming all exporter-year, importer-year, and border-year fixed effects. In words, the potential outcomes prior to treatment are, on average and conditional on the fixed effects, the same, i.e., there is no effect of the treatment before its onset. As discussed, for example, in Callaway and Sant'Anna (2021) and Wooldridge (2023), in principle, the no anticipation assumption does not need to hold in all pre-treatment periods and can be relaxed by omitting conspicuous periods at the cost of efficiency.

Second, the parallel trends assumption requires for all $t \in \{1, \ldots, T\}$ that

$$\frac{E[Y_t(\infty)|D_q, \ldots, D_T, \mathbf{X}]}{E[Y_1(\infty)|D_q, \ldots, D_T, \mathbf{X}]} = \exp\left[\mathbf{X}\pi_t\right], \quad (6)$$

where $\pi_\mathbf{t}$ captures all the coefficients corresponding to exporter-year, importer-year, and border-year fixed effects. In words, the parallel trends assumption states that the growth in the outcome of the treatment and control group would have been the same in the

[16]Note that equation (3) differs from the setting considered in Wooldridge (2023) with regard to its additional fixed effect structure (against the background of the more complex double-indexed panel data under consideration).

absence of treatment, i.e., growth may depend on the time-varying covariates $\mathbf{X}$, but not on the treatment $\mathbf{D_i} = (D_{iq}, \ldots, D_{iT})$. Note that due to the use of the exponential mean in equation (3), the parallel trends assumption is stated in terms of the ratio of means rather than the difference in means as in the linear case (see also, e.g., Athey and Imbens, 2006; Ciani and Fisher, 2019).

Lastly, two additional points are worth highlighting. First, just like the TWFE estimator, the ETWFE identifies partial effects of RTAs on trade due to the inclusion of exporter-year and importer-year fixed effects, which capture multilateral resistance terms and which are expected to change as a result of RTAs. The total effect of RTAs on trade, including these general equilibrium effects could, in principle, be backed out using computational approaches that are now standard in the trade literature and can even be performed with the PPML estimator directly in Stata, e.g., as described in Anderson et al. (2018). Second, note that there is a certain asymmetry between the restrictions on non-treated potential outcomes and restrictions on the treatment effects (see footnote 9 and page 10 of Borusyak et al. (2023)). Here, we follow the literature on heterogeneity-robust DiD estimators, which takes the view that restrictions on the form of non-treated potential outcomes are acceptable, while homogeneity restrictions on treatment effects are not, which is not a limitation that is specific to our paper.

### 2.2.2 Estimation target

The ETWFE estimator proposed in Wooldridge (2021, 2023) initially requires estimation of treatment effects at a very granular level, which is a common feature in many heterogeneity-robust DiD estimators proposed in the recent literature (e.g. Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2023; de Chaisemartin and D'Haultfoeuille, 2022). However, interest often lies in more aggregated quantities rather than the directly estimated cohort-time-specific treatment effects from equation (3). Consistent with the large corresponding trade literature, in this paper, we are interested in evaluating by how much international trade has grown, on average, in all country pairs

and post-intervention years as a result of the introduction of RTAs. Therefore, our main estimation target is a single RTA estimate, which is defined by the weighted sum of the estimated cohort-time-specific treatment effects as:

$$\overline{\hat{\delta}} = \sum_{g=q}^{T} \sum_{s=g}^{T} \frac{N_{gs}}{N_D} \hat{\delta}_{gs}, \tag{7}$$

where we assign equal weight to all post-treatment observations, which correspond to the number of observations of cohort $g$ in period $s$, $N_{gs}$, relative to the total number of treated observations, $N_D = \sum_{g=q}^{T} \sum_{s=g}^{T} N_{gs}$. Similarly, the standard errors for the ETWFE estimator are computed as a (weighted) linear combination of the cohort-year-specific effects taking the covariance between the coefficients into account. The reasons for the selection of these weights are that we view them as the simplest, most intuitive, and most transparent option, which is also consistent with de Chaisemartin and D'Haultfœuille (2020) and Borusyak et al. (2023).[17] Naturally, in other settings and for other research questions, different weighting schemes may be more appropriate.[18]

Note that, since $\hat{\delta}_{gs}$ is defined by a log difference (cf. equation (4)), equation (7) has the attractive property that it is directly related to the geometric mean, as defined by the arithmetic mean in logscale. Therefore, $\exp(\hat{\delta}) - 1$ computes the geometric growth rate of the RTA effect on trade, which takes account of compounding and is more appropriate than the arithmetic mean in our setting with exponential growth.

We also consider two additional estimation targets. First, we consider a cohort estimand, for which we compute cohort-specific treatment effects by averaging over the time dimension as follows:

$$\overline{\hat{\delta}_{g\cdot}} = \sum_{s=g}^{T} \frac{N_{gs}}{N_{g\cdot}} \hat{\delta}_{gs}, \tag{8}$$

where $N_{g\cdot} = \sum_{s=g}^{T} N_{gs}$ is the total number of post-treatment observations of cohort $g$.

[17] In the robustness analysis in Section 5, we also consider the robustness of our results to alternative weighting schemes.

[18] See also the discussion on this point in Borusyak et al. (2023).

Second, we consider an event-study estimand, for which we compute event-time-specific treatment effects by averaging over the cohort dimension as follows:

$$\overline{\hat{\delta}_{\cdot s}} = \sum_{g=q}^{s} \frac{N_{gs}}{N_{\cdot s}} \hat{\delta}_{gs}, \tag{9}$$

where $N_{\cdot s} = \sum_{g=q}^{s} N_{gs}$ is the total number of treated observations in period $s$. Event-time-specific treatment effects are akin to results from a dynamic TWFE (or event-study) specification without its shortcomings based on the heterogeneity-robust innovations in the difference-in-differences (DiD) literature (e.g., Sun and Abraham, 2021; Borusyak et al., 2023), while cohort-specific treatment effects are usually not studied in the associated TWFE literature.

As discussed in detail in Callaway and Sant'Anna (2021), caution needs to be exercised when comparing $\overline{\hat{\delta}_{g\cdot}}$ across cohorts or $\overline{\hat{\delta}_{\cdot s}}$ across time periods since compositional changes due to differences in weights and in the composition of groups/periods can complicate their interpretation. For example, early-treated cohorts have more observations many years after the treatment onset than late-treated cohorts. Consequently, in the presence of treatment effect heterogeneity across cohorts, changes in event-time-specific treatment effects are the result of changes in event time as well as the higher share of early-treated cohorts relative to late-treated cohorts. Similarly, in the presence of treatment effect heterogeneity across event time, differences in cohort-specific treatment effects are due to differences across cohorts as well as the higher share of long-term effects in early-treated cohorts relative to late-treated cohorts.

### 2.2.3 On the bias of the TWFE estimator: An illustrating example

This subsection provides a small stylized example to illustrate the potential bias of the TWFE estimator in the presence of treatment effect heterogeneity. Here, we focus on the OLS TWFE estimator, which can be shown to identify some weighted average of the true underlying heterogeneous cohort-year effects under standard identifying assumptions

(Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020), and we consider the example studied in de Chaisemartin and D'Haultfœuille (2020) and Borusyak et al. (2023) with two equal-sized cohorts (or units) and three time periods.

The early-treated cohort A is treated in period 2 and the late-treated cohort B is treated in period 3. The treatment effect for cohort A in period 2 and 3 is denoted by $\delta_{A2}$ and $\delta_{A3}$, respectively, and the treatment effect for cohort B in period 3 is denoted by $\delta_{B3}$. The outcome of cohort $i$ in period $t$ is denoted by $Y_{it}$. The TWFE estimand can be shown to equal $\delta_{TWFE} = \delta_{A2} - \frac{1}{2}\delta_{A3} + \frac{1}{2}\delta_{B3}$ under a parallel trends and a no anticipation assumption (Borusyak et al., 2023). By contrast, the corresponding ETWFE estimand from Section 2.2.2 would be $\delta_{ETWFE} = \frac{1}{3}\delta_{A2} + \frac{1}{3}\delta_{A3} + \frac{1}{3}\delta_{B3}$, i.e., the weight equals the proportion of each cohort-year cell in all post-treatment observations.

This example illustrates several points. First, note that if the assumption of treatment effect homogeneity of the TWFE estimator is fulfilled, then the TWFE estimand correctly identifies the true underlying treatment effect ($\delta_{TWFE} = \delta_{A2} = \delta_{A3} = \delta_{B3}$). Second, the problem of the TWFE estimator in the presence of treatment effect heterogeneity results from 'forbidden comparisons' (Borusyak and Jaravel, 2017). An 'admissible comparison' would be to compare the evolution of cohort A and B between periods 2 to 1, $((Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1}))$. By contrast, in this example, the TWFE estimator also uses a 'forbidden comparison', in which the evolution of cohort B and A between period 3 to 2, $((Y_{B3} - Y_{B2}) - (Y_{A3} - Y_{A2}))$, is used. However, subtracting the term $(Y_{A3} - Y_{A2})$, inter alia, also subtracts the evolution of treatment effects, $\delta_{A3}$ - $\delta_{A2}$, thereby placing a negative weight on the late treatment effect of the early-treated cohort A (Borusyak et al., 2023).

Third, in comparison to the ETWFE estimand, the TWFE estimand underweights early-treated cohorts and overweights late-treated cohorts. For example, in the case considered above, the weight placed on cohort A (B) is $\frac{1}{2}$ ($\frac{1}{2}$) for the TWFE estimand and $\frac{2}{3}$ ($\frac{1}{3}$) for the ETWFE estimand. If treatment effects decrease (increase) over cohorts, this results in a downward (upward) bias of the TWFE estimand. As demonstrated in the empirical analysis below, this is exactly the case for the effects of regional trade agreements,

thus highlighting the benefits of the ETWFE estimator.

Fourth, in comparison to the ETWFE estimand, the TWFE estimand overweights short-run effects and underweights long-run effects resulting in a severe short-run bias of the static TWFE specification (Borusyak et al., 2023). For example, in the case considered above, the weight placed on the first (second) year of treatment is $1\frac{1}{2}$ $(-\frac{1}{2})$ for the TWFE estimand and $\frac{2}{3}$ $(\frac{1}{3})$ for the ETWFE estimand. If treatment effects increase (decrease) over time, this also results in a downward (upward) bias of the TWFE estimand. Like above, we also find empirical support for this phenomenon for the effects of regional trade agreements, which are typically characterized by strong maturation effects.

Fifth, de Chaisemartin and D'Haultfœuille (2020) generalize the results in points three and four. They show that negative weights are more likely to occur in time periods when a large fraction of groups are treated and in groups treated for many periods (i.e., late treatment effect of the early-treated cohorts).

Lastly, note that the extent of negative weights depends on a number of factors and that they might entirely disappear with a large number of never-treated and not-yet treated observations (Borusyak et al., 2023). However, even in this case, the weights of the TWFE estimand might strongly diverge from those of the ETWFE estimand in the direction described above.[19]

# 3 Data

Given the methodological purpose of our paper, for our empirical analysis, we rely on three standard datasets, including data on (i) trade flows, (ii) RTAs, and (iii) other gravity variables. The trade data that we use to obtain our estimates are from the Structural Gravity Database (SGD) of the World Trade Organization (Larch et al., 2019a).[20] The SGD contains aggregate manufacturing trade data and has three important advantages for

[19] We offer an analysis of the underlying weights of the OLS TWFE estimator in our setting in Section 4.4.

[20] The SGD is available at https://www.wto.org/english/res_e/reser_e/structural_gravity_e.htm.

our purposes. First, it covers a long period of time, 1980-2016. Second, it contains many countries (a total of 229). In combination, this implies that we could have 'long T' and 'long N' panel datasets for our analysis, and we experiment with alternative sample sizes.[21] Third, the SGD includes consistently constructed domestic and international trade flows, and we demonstrate that our methods have implications for gravity estimations with and without domestic trade flows.[22]

In addition to the SGD, we use the Dynamic Gravity Dataset (DGD) of the United States International Trade Commission (Gurevich and Herman, 2018) for data on trade agreements and some other standard gravity variables (e.g., bilateral distance, contiguity, WTO membership, etc.) that we employ in the robustness analysis.[23] This database covers all trade agreements in the world that entered into force during the period for which trade data are available.[24] As briefly discussed in the introduction, the focus on RTAs is suitable for our purposes for several reasons. First, the impact of RTAs is probably the most studied topic from a trade policy perspective. In addition, there is ample empirical evidence that the effects of RTAs are heterogeneous across groups, e.g., agreements and pairs within agreements (Baier et al., 2019). Thus, existing RTA estimates may indeed be subject to the recent critiques from the staggered DiD literature.

A known challenge with the existing RTA datasets is that they report the date of the effective entry into force of the RTA as the initial RTA date. However, in reality, the onset of the effects of RTAs usually does not coincide with the date of their entry into force. The reason is that once an RTA is announced, which usually happens several years prior to its effective implementation, some trade barriers may start falling and/or some economic agents may start adjusting in anticipation of the RTA (e.g., Moser and

[21] To obtain our main results, we use a sample of 69 countries, which covers 98% of world exports. We also experiment with a medium sample, which contains 91 countries covering 99% of world exports, and also with a large sample, which covers all the countries from the SGD. Due to singletons, the actual number of countries in our large estimating sample is 225 (out of 229).

[22] The use of domestic trade flows, in addition to international trade, is consistent with gravity theory and has a number of potential advantages for gravity estimations. Yotov (2022) reviews the literature.

[23] The DGD is freely available at https://www.usitc.gov/data/gravity/dgd.htm.

[24] Other excellent RTA dataset include, for example, the Regional Trade Agreements Database of Egger and Larch (2008) and the NSF-Kellogg Institute Data Base on Economic Integration Agreements of Baier and Bergstrand (2021).

Rose, 2012a, 2014; Egger et al., 2022). Thus, arguably, the announcement of the RTA negotiations is probably the more appropriate onset for studying the impact of RTAs.

Unfortunately, as noted earlier, we are not aware of a comprehensive dataset on the announcement of RTAs. However, there is some evidence that could guide us. For example, according to Moser and Rose (2012b) the period between the start of the initial negotiations and the signing of a trade agreement is about 28 months. In addition, according to the WTO data, the average period between the signing and the entry into force of an agreement is a about one year. In combination, this implies that, on average, the RTA negotiations start about three years before the date of entry into force. This is consistent with the results from Egger et al. (2022), who find that the impact of RTAs begins about three years prior to their entry into force, possibly at the time when they are announced or signed.

Based on this, for our main analysis, we will assume an 'onset' of RTAs three years before their entry into force.[25] As a robustness test, following Wooldridge (2023), we also omit these time periods in Section 5. For econometric reasons, we exclude country pairs with an RTA onset prior to 1980 (so-called always-treated units), since the RTA effect on trade cannot be identified in these cases given the absence of pre-treatment periods (and since they also cannot serve as control observations in the presence of treatment effect heterogeneity).

Table 1 reports the number of observations, country pairs, exporters, importers, and years for different groups in the data set. A cohort refers to all country pairs with an RTA onset in a particular year. First, Table 1 highlights that the data set contains a reasonable number of pre- and post-treatment observations for most cohorts. Second, Table 1 captures the sharp increase in RTAs in the late 1980s and the 1990s. Another feature of the evolution of the RTAs, which is not obvious from Table 1, is that trade agreements have become increasingly 'deeper' over time, i.e., they contain more provisions for trade liberalization. This type of heterogeneity could further motivate the need for

[25]Note that we only consider RTAs that entered into force in the time period until 2016 in our sample.

the proposed ETWFE estimator.

Table 1: Descriptive statistics: Observations along different dimensions

| Group | Observations | Pairs | Exporters | Importers | Years |
|---|---|---|---|---|---|
| 1985 cohort | 64 | 2 | 2 | 2 | 32 |
| 1986 cohort | 620 | 20 | 9 | 9 | 31 |
| 1989 cohort | 6,927 | 251 | 26 | 26 | 28 |
| 1990 cohort | 375 | 15 | 9 | 9 | 27 |
| 1991 cohort | 385 | 15 | 9 | 9 | 26 |
| 1992 cohort | 1,066 | 43 | 15 | 15 | 25 |
| 1993 cohort | 1,029 | 43 | 19 | 19 | 24 |
| 1994 cohort | 573 | 25 | 13 | 13 | 23 |
| 1995 cohort | 592 | 27 | 9 | 9 | 22 |
| 1996 cohort | 81 | 4 | 4 | 4 | 21 |
| 1997 cohort | 480 | 24 | 12 | 12 | 20 |
| 1998 cohort | 847 | 45 | 18 | 18 | 19 |
| 1999 cohort | 216 | 12 | 8 | 8 | 18 |
| 2000 cohort | 918 | 54 | 20 | 20 | 17 |
| 2001 cohort | 224 | 14 | 11 | 11 | 16 |
| 2002 cohort | 570 | 38 | 22 | 22 | 15 |
| 2003 cohort | 560 | 40 | 24 | 24 | 14 |
| 2004 cohort | 2,047 | 158 | 41 | 41 | 13 |
| 2005 cohort | 144 | 12 | 9 | 9 | 12 |
| 2006 cohort | 220 | 20 | 13 | 13 | 11 |
| 2007 cohort | 259 | 26 | 15 | 15 | 10 |
| 2008 cohort | 252 | 30 | 18 | 18 | 9 |
| 2009 cohort | 160 | 20 | 13 | 13 | 8 |
| 2010 cohort | 126 | 18 | 9 | 9 | 7 |
| 2011 cohort | 288 | 48 | 29 | 29 | 6 |
| 2012 cohort | 100 | 20 | 13 | 13 | 5 |
| 2013 cohort | 431 | 108 | 39 | 39 | 4 |
| 2014 cohort | 18 | 6 | 6 | 6 | 3 |
| 2015 cohort | 20 | 10 | 8 | 8 | 2 |
| 2016 cohort | 50 | 50 | 28 | 28 | 1 |
| Treated | 19,642 | 1,198 | 66 | 66 | 32 |
| Not-yet treated | 15,951 | 1,198 | 66 | 66 | 33 |
| Never treated | 69,816 | 2,599 | 69 | 69 | 34 |

Third, to state the obvious, early-treated cohorts have more post-treatment years than late-treated cohorts, implying that longer-run effects can only be estimated for the former. Fourth, note that the data set contains two control groups, never-treated country pairs and non-yet-treated country pairs, that can be used for comparisons with the post-treatment years of the treatment group. Never-treated country pairs are those that did not sign an agreement in the time period of our sample. Not-yet-treated country pairs are those without an RTA onset up until the year of the comparison, but with an RTA onset in later years of the sample.

Table 2 reports the average of the variables Distance (in kilometers), Contiguity, Lan-

guage, and Colony for different groups in the baseline estimation sample from the ETWFE estimate in column (2) of Table 3. 'Cohort' refers to all country pairs with an RTA onset in a particular year. 'Treated' refers to all cohorts. 'Never treated' refers to all country pairs that did not sign an RTA agreement during the sample period.

Table 2: Descriptive statistics: Summary statistics of covariates for different groups

| Group | Distance | Contiguity | Language | Colony |
|---|---|---|---|---|
| 1985 cohort | 10,512 | 0.00 | 1.00 | 0.00 |
| 1986 cohort | 2,235 | 0.00 | 0.00 | 0.10 |
| 1989 cohort | 10,607 | 0.02 | 0.49 | 0.01 |
| 1990 cohort | 9,635 | 0.00 | 0.41 | 0.00 |
| 1991 cohort | 5,264 | 0.00 | 0.43 | 0.00 |
| 1992 cohort | 1,577 | 0.07 | 0.04 | 0.00 |
| 1993 cohort | 2,238 | 0.00 | 0.14 | 0.00 |
| 1994 cohort | 2,419 | 0.17 | 0.17 | 0.00 |
| 1995 cohort | 2,843 | 0.00 | 0.08 | 0.00 |
| 1996 cohort | 6,317 | 0.00 | 0.00 | 0.00 |
| 1997 cohort | 3,016 | 0.16 | 0.46 | 0.00 |
| 1998 cohort | 3,043 | 0.03 | 0.64 | 0.00 |
| 1999 cohort | 1,477 | 0.30 | 0.49 | 0.16 |
| 2000 cohort | 8,460 | 0.00 | 0.15 | 0.04 |
| 2001 cohort | 5,543 | 0.15 | 0.31 | 0.00 |
| 2002 cohort | 1,647 | 0.00 | 0.25 | 0.00 |
| 2003 cohort | 9,180 | 0.06 | 0.59 | 0.03 |
| 2004 cohort | 5,055 | 0.03 | 0.12 | 0.01 |
| 2005 cohort | 7,767 | 0.00 | 0.67 | 0.00 |
| 2006 cohort | 9,097 | 0.00 | 0.40 | 0.00 |
| 2007 cohort | 5,992 | 0.08 | 0.09 | 0.00 |
| 2008 cohort | 7,515 | 0.00 | 0.08 | 0.00 |
| 2009 cohort | 6,581 | 0.00 | 0.68 | 0.00 |
| 2010 cohort | 9,221 | 0.00 | 0.78 | 0.00 |
| 2011 cohort | 8,409 | 0.00 | 0.27 | 0.00 |
| 2012 cohort | 8,200 | 0.00 | 0.64 | 0.00 |
| 2013 cohort | 9,087 | 0.00 | 0.10 | 0.02 |
| 2014 cohort | 8,556 | 0.00 | 0.33 | 0.00 |
| 2015 cohort | 8,749 | 0.00 | 0.60 | 0.00 |
| 2016 cohort | 2,880 | 0.15 | 0.39 | 0.00 |
| Treated | 6,885 | 0.03 | 0.32 | 0.01 |
| Never treated | 8,286 | 0.01 | 0.26 | 0.02 |

We note that distance, in particular, varies strongly across different treatment cohorts. In Section 5, we exploit this variation in an additional specification akin to a regression adjustment approach (Heckman et al., 1997; Wooldridge, 2023) that relaxes the parallel trend assumption of the ETWFE estimator.

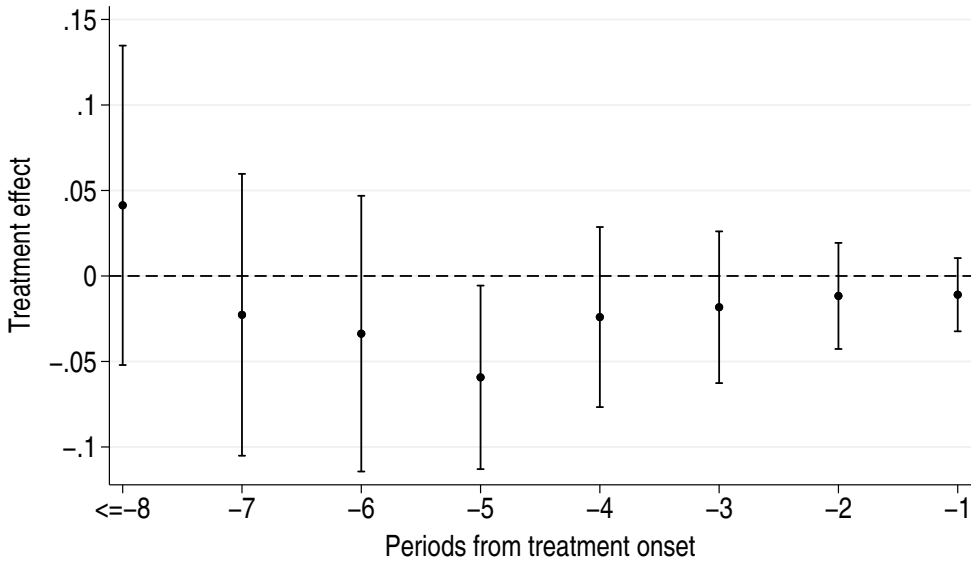
# 4 Main results and analysis

This section presents our main findings. We develop the analysis in four steps. First, we start by testing the identifying assumptions of the ETWFE estimator (Section 4.1). Second, we obtain and discuss our main results (Section 4.2). Third, we present event-study type estimates that describe the evolution of the RTA effects over time, as well as a series of cohort-specific RTA estimates (Section 4.3). Fourth, we compute the implicit weights underlying the (OLS) TWFE estimator (Section 4.4).

## 4.1 Test of identifying assumptions

The ETWFE estimator requires the parallel trends assumptions and the no anticipation assumption (described in detail in Section 2.2.1) to hold. Therefore, we begin by testing these two assumptions in Figure 2.

Figure 2: Pre-treatment effects




**Notes**: The figure reports pre-trend estimates from a PPML estimation of equation (3), in which treatment effects were replaced by cohort-time-specific placebo treatment effects prior to treatment onset. The regression is estimated with untreated observations only in the spirit of Borusyak et al. (2023). The cohort-time-specific treatment effects were aggregated using equation (9) to obtain event-time-specific treatment effect estimates. 95% confidence intervals are shown using standard errors clustered by country pair.

To this end, we adapt to our setting the methods of Borusyak et al. (2023).[26] Specifically, we replace the treatment effects in equation (3) with cohort-time-specific placebo treatment effects prior to treatment onset and estimate the regression with untreated observations only. Figure 2 provides evidence against the violation of the identifying assumptions by showing the aggregated placebo effects in event time relative to treatment onset. The pre-treatment coefficient are all close to zero and a joint hypothesis test is insignificant (p-value of 0.2978). Therefore, we conclude that there are no signs of significant pre-trends in our setting and we proceed to obtain our main results.

## 4.2 Main estimation results

Our main estimates are reported in Table 3. The RTA estimate in column (1) corresponds to the 'Consecutive Years: 2022' result from Figure 1 and is obtained with all current techniques from the gravity literature, including the PPML estimator, exporter-time and importer-time fixed effects, pair fixed effects, domestic trade flows, globalization effects, consecutive-year data, and pair-clustered standard errors as specified in equation (1). The RTA estimate is positive and statistically significant (0.166, std.err. 0.050), implying an increase in the bilateral trade between member countries of about 18% (calculated as $(\exp(0.166)-1)\times 100 = 18.067$, std.err. 5.913), where the standard error is obtained with the Delta method.

Our ETWFE RTA estimate appears in column (2) of Table 3, and it is obtained from equation (3), i.e., the same specification from column (1), the only difference being that, in addition to all gravity estimation techniques, we also implement a heterogeneity-robust staggered DiD method. Note that the cohort-time-specific treatment effects were aggregated using equation (7) to obtain the aggregate treatment effect estimate. The aggregate ETWFE estimate is positive and statistically significant, as expected.[27] However, eco-

[26] Alternatively, using the parallel trend test based on the entire sample discussed in Wooldridge (2023) yields similar results. We note that in this setting the two tests are not equivalent due to the more complex fixed effect structure under consideration.

[27] See Section 5 for a discussion of the standard error estimate.

nomically, it is substantially larger than the corresponding TWFE result from column (1), implying that the RTAs in our sample have led to an increase of about 46% in trade among member countries. A standard F-test confirms that the two coefficients are also statistically significantly different from each other.

Table 3: Comparison of TWFE and ETWFE estimators for different samples

| | (1) TWFE | (2) ETWFE | (3) TWFE | (4) ETWFE | (5) TWFE | (6) ETWFE |
|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.166*** | 0.381*** | 0.167*** | 0.327*** | 0.165*** | 0.293*** |
| | (0.050) | (0.041) | (0.048) | (0.040) | (0.047) | (0.039) |
| Sample | Baseline | Baseline | Medium | Medium | Large | Large |
| Observations | 105,409 | 105,409 | 175,796 | 175,796 | 591,092 | 591,092 |
| Exporters | 69 | 69 | 91 | 91 | 225 | 225 |
| Importers | 69 | 69 | 91 | 91 | 225 | 225 |
| Years | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 1 | 469 | 1 | 469 | 1 | 528 |
| p-value H0:TWFE=ETWFE | | 0.000 | | 0.000 | | 0.001 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Importer × year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Cross-border × year FE | Yes | Yes | Yes | Yes | Yes | Yes |

**Notes**: The table presents PPML regression results using the TWFE estimator (equation (1)) and the ETWFE estimator (equation (3)), for which the cohort-time-specific treatment effects were aggregated using equation (7) to obtain an aggregate treatment effect estimate. The dependent variable is exports, which vary over the exporter-importer-year dimension. The 'Baseline' sample contains 69 countries, accounting for 98% of world exports. The 'Medium' sample, which is used to obtain the results in columns (3) and (4), contains 91 countries, accounting for 99% of world exports. The 'Large' sample, which is used to obtain the results in columns (5) and (6), contains the full set of countries from the structural gravity dataset. 'Coefficients' reports the number of estimated coefficients apart from the fixed effects. Standard errors in parentheses are clustered by country pair. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. The row labeled "p-value H0: TWFE=ETWFE" reports the results of a standard F-test of the equality of the corresponding TWFE and ETWFE coefficients, where the covariance between the two coefficients is accounted for using a seemingly unrelated regression specification.

Our main result has two implications. First, from a methodological perspective, it reveals that there is still scope for meaningful improvements in the techniques that are used to identify the impact of RTAs within gravity models. Second, the fact that the ETWFE estimate is significantly larger is important from a policy perspective since, as discussed earlier, some of the existing RTA estimates are viewed as unrealistically small.

The rest of the results in Table 3 are obtained with two alternative samples. Specifically, the estimates in columns (3) and (4) are based on a sample that contains 91 countries, accounting for 99% of world exports, while the estimates in columns (5) and (6) are obtained with the full set of countries from the WTO dataset. We draw the following

conclusions based on comparisons between the estimates that we obtain with the alternative samples. The TWFE estimate remains very stable, while the ETWFE estimate decreases with the increase of the number of countries in our sample. As a result, the gap between the TWFE and ETWFE estimates becomes smaller; however, the ETWFE estimate is still about twice as large as the corresponding TWFE estimate.

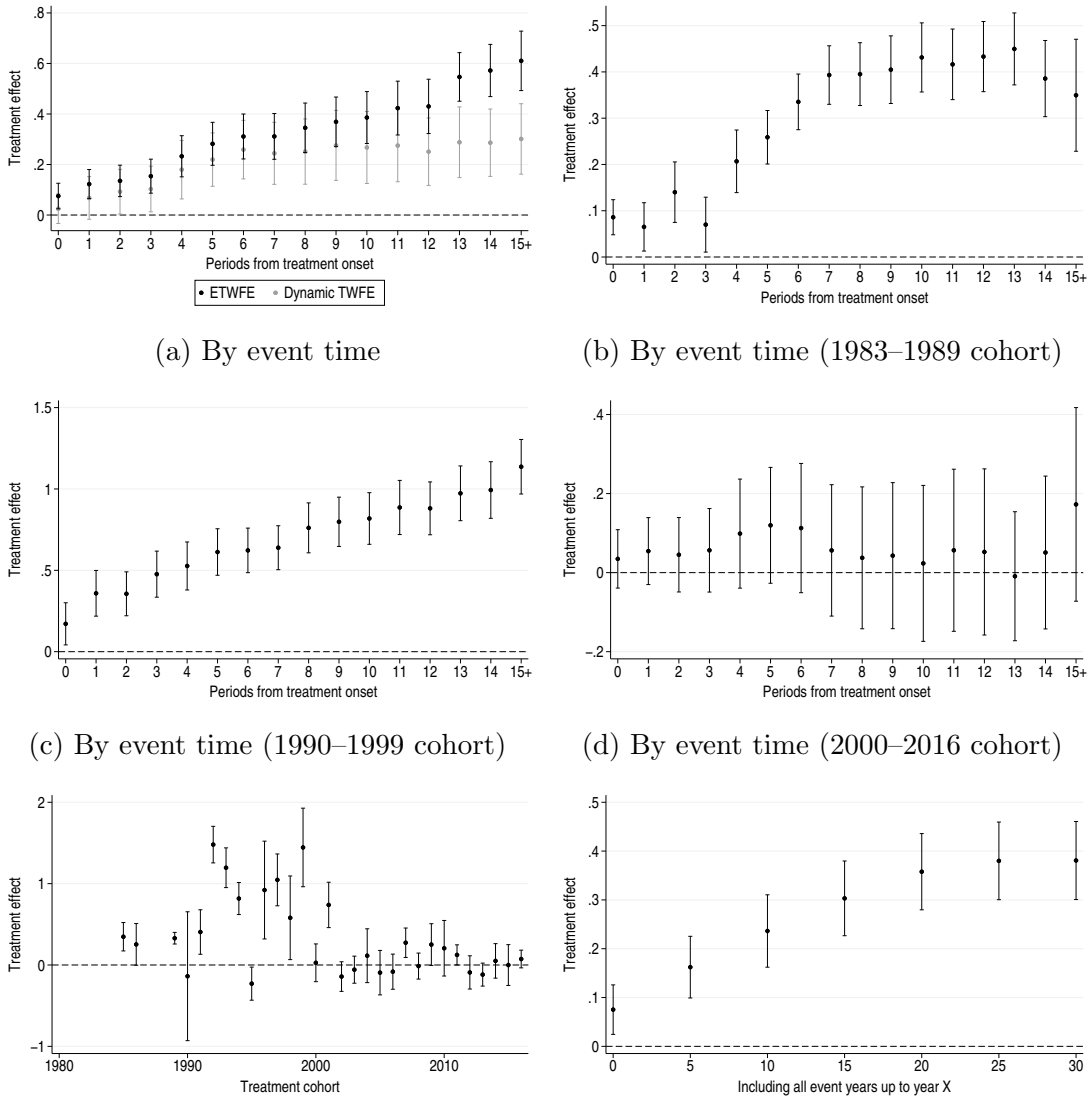
To investigate the source of the difference in the ETWFE estimates across the different samples, we compute separate treatment effects for the treated country pairs in the original sample and those that are additionally included in the medium and large sample. The treatment effect for the original treatment group remains virtually unchanged (medium sample: 0.376, std.err. 0.039; large sample: 0.387, std.err. 0.038), while the treatment effect of the new country pairs is substantially smaller (medium sample: 0.236, std.err. 0.046; large sample: 0.244, std.err. 0.041). This suggests that the differences are driven by smaller treatment effects in the additional country pairs rather than changes in the composition of the control group. From a methodological perspective, the TWFE estimator might miss these changes for the following reason. While a larger share of smaller treatment effects are expected to reduce the TWFE estimate, the use of these already-treated country pairs with smaller treatment effects in the control group is also expected to raise the overall RTA estimate. Taken together, these two effects appear to offset each other such that the TWFE estimator misses the lower RTA effects in the larger samples, which underscores the importance of using the ETWFE estimator.

## 4.3 Disaggregated results

We complement the average RTA estimates from Table 3 with a series of disaggregated results. We start with an event-type analysis of the evolution of the RTA effects over time. Our findings are presented in Figure 3a, where we report standard event-study (dynamic TWFE) results in light color and the corresponding ETWFE results in dark color, which are computed from the cohort-time-specific treatment effects by averaging over the cohort

dimension (see Section 2.2.2).[28]

Figure 3: Event-time-specific and cohort-specific treatment effects




(a) By event time

(b) By event time (1983–1989 cohort)

(c) By event time (1990–1999 cohort)

(d) By event time (2000–2016 cohort)

(e) By cohorts

(f) By event-window length

**Notes**: The figure reports different aggregations of cohort-time-specific treatment effects from PPML estimation of equation (3). Panel (a) reports event-time-specific treatment effects from equation (3) aggregated using equation (9) in dark color ('ETWFE') along with standard event-study results in light color ('dynamic TWFE'). Panels (b)–(d) report event-time-specific treatment effects from equation (3) aggregated using equation (9) for the 1983-1989 cohort, the 1990-1999 cohort, and the 2000-2016 cohort, respectively. Panel (e) reports cohort-specific treatment effects from equation (3) aggregated using equation (8). Panel (f) reports treatment effects aggregated using equation (7) in which only those cohort-time-specific treatment effects are used in the aggregation that are not more than a certain number of years away from treatment onset. 95% confidence intervals are shown using standard errors clustered by country pair.

Two main results stand out from Figure 3a. First, the TWFE estimates are smaller in each period. This is similar to the smaller average RTA effects from the static TWFE

[28]The event-study (dynamic TWFE) estimator corresponds to a restricted version of the ETWFE estimator, in which treatment effect homogeneity across cohorts is imposed (cf. column (1) in Table 4).

estimator from Table 3 and, therefore, reinforces our main result. Second, according to our dynamic TWFE estimates, the impact of RTAs is exhausted relatively fast, e.g., about 6 years after their initial effects are detected.[29] The ETWFE estimates, on the other hand, appear to keep growing slowly, but monotonically, more than 15 years after the initial impact of the RTAs. This is likely another reason for the difference between the average estimates from columns (1) and (2) of Table 3. In addition, from a policy perspective, the estimates from Figure 3a imply that the effects of the RTAs in our sample have lasted significantly longer than suggested by the dynamic TWFE estimates.

Next, we capitalize on the advantages of the ETWFE estimator to obtain additional estimates of the effects of RTAs across different dimensions. Figure 3b–d reports event-study type results for three different groups of cohorts: the 1983-1989 cohort, the 1990-1999 cohort, and the 2000-2016 cohort, where, e.g., the 1983-1989 cohort refers to all country pairs with an RTA onset between 1983 and 1989. They show that there is substantial heterogeneity across cohorts. For the 1983-1989 cohort, the RTA effect increases and then saturates around 10 time periods after treatment onset, while for the 1990-1999 cohort the RTA effect seems to steadily increase even after 15 years. By contrast, there is no significant effect in any time period for the 2000-2016 cohort.

Figure 3e zooms further in on cohort heterogeneity. It shows average cohort-specific treatment effects, which are computed from the cohort-time-specific treatment effects by averaging over the time dimension (see Section 2.2.2). In line with the previous results, there is substantial heterogeneity across treatment cohorts. RTAs with an onset in the 1980s have positive effects, but mostly lower than the average RTA estimates from column (2) in Table 3. The effects of RTAs with an onset in the 1990s are largest and mostly above average, while the cohorts after 2000, on average, do not show significant effects.

We have several intuitive explanations for the heterogeneity of our cohort-specific RTA estimates. First, the large estimates for the cohort of agreements in the 1990s are

[29]Note that changes in the estimate over time periods are also affected by changes in the composition of the underlying cohorts since, e.g., RTAs signed in later years of the sample are observed for fewer years than RTAs signed in earlier years of the sample.

consistent with the perception of this period as the 'golden age' of trade liberalization. Second, on a related note, the most 'natural' RTAs were already signed during the 1980s and 1990s and, therefore, the potential of the more recent agreements, e.g., those in the 2000s, to lead to significant increases in trade among members was lower and more limited. Third, the first two decades of the 2000s was marked by significant economic crises (e.g., the 'dot.com' bust and the financial crises), followed by Brexit and the Trump presidency, which all have impacted trade liberalization efforts negatively. Fourth, average applied most-favored nation tariffs fell significantly over the sample period (e.g., Teti, 2020), thereby leaving less scope for trade-enhancing effects of RTA-related tariff reductions.[30]

Lastly, we look at the role that the event-window length and hence years far away from treatment onset play for the magnitude of the aggregate effect. Figure 3f reports average treatment effects in which only those cohort-time-specific treatment effects are used in the aggregation that are not more than a certain number of years away from treatment onset. In our sample, the aggregate treatment effect steadily increases with event-window length at least up until 20 years and then levels off. This suggests that truncating the sample may bias the aggregate treatment effect estimate downward given the long-lasting dynamic effects of RTAs (e.g., Schmidheiny and Siegloch, 2020; Baker et al., 2022).

We draw two methodological and two policy conclusions based on the analysis in this section. From a methodological perspective, our analysis demonstrates that the ETWFE RTA effects are (i) significantly larger than previously thought based on TWFE estimates, and (ii) that the RTA effects last significantly longer than suggested by the dynamic TWFE estimates. From a policy perspective, our analysis reveals that (i) the effects of RTAs phase in gradually and (ii) that recent trade agreements have not been very successful in promoting international trade.

[30] We also recognize that RTAs have become deeper and more comprehensive over time. This may have countervailing effects with respect to the impact of RTAs. On the one hand, deeper trade liberalization should lead to more trade among members. On the other hand, more cumbersome and complicated RTA provisions may impede trade. We explore these hypotheses further in the robustness section Section 5.

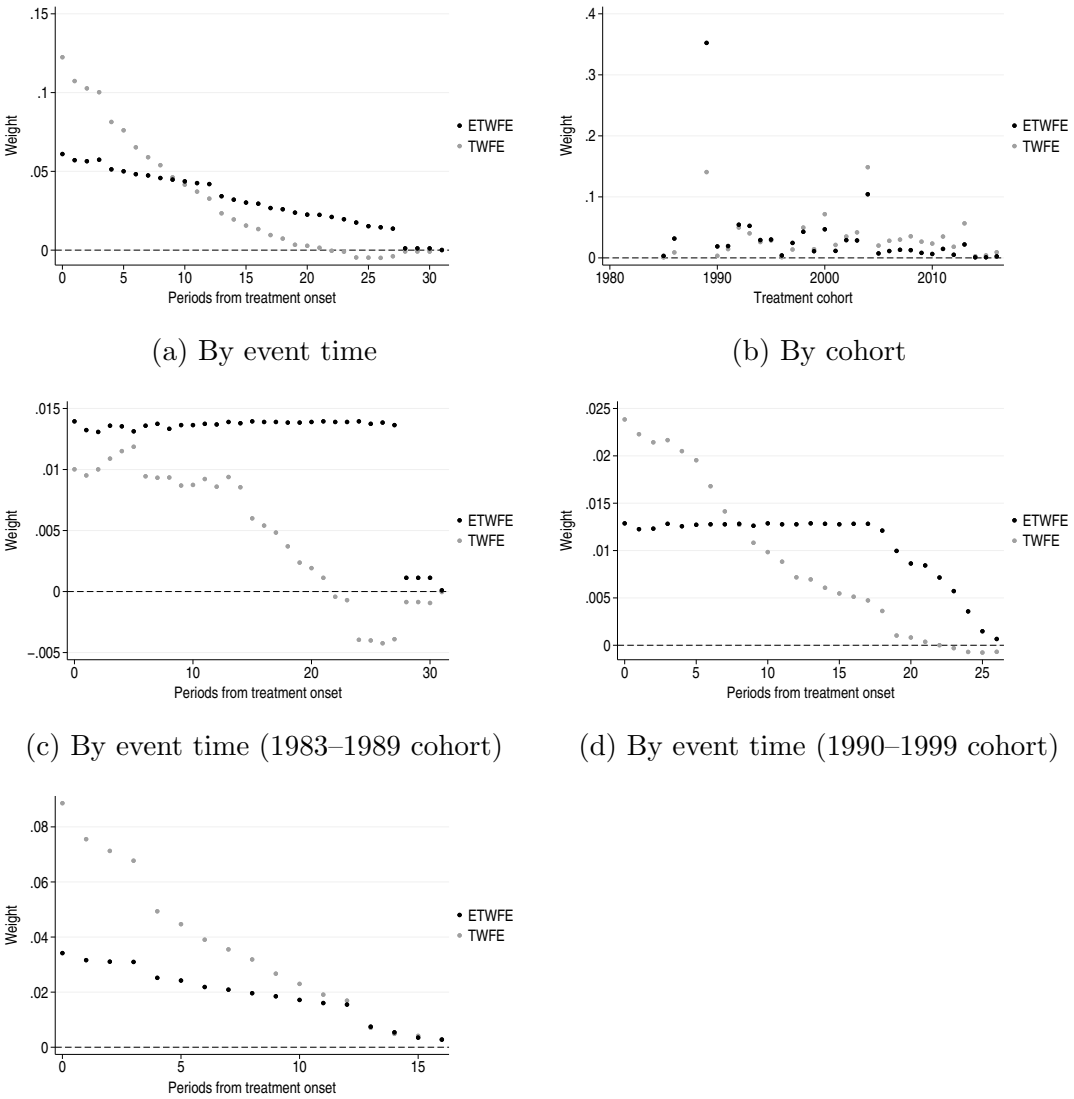
## 4.4 Implicit weights attached by the OLS TWFE estimator

In this section, we follow de Chaisemartin and D'Haultfœuille (2020) to compute the implicit weights attached by the OLS TWFE estimator to individual cohort-year cells. We proceed in two steps. First, we reproduce our main results from columns (1) and (2) of Table 3 with the OLS estimator to ensure that they are comparable to the PPML results. For the baseline sample, the OLS TWFE point estimate (0.172) is very close to the corresponding PPML result (0.166), while the OLS ETWFE point estimate (0.347) is only slightly smaller than the corresponding PPML result (0.381). See also Section 5.4 of the robustness analysis for details.

Second, we follow de Chaisemartin and D'Haultfœuille (2020) to decompose the OLS TWFE estimate into cohort-year specific treatment effects (from the OLS ETWFE estimation) and implicit weights attached to these cohort-year cells. By construction, the corresponding weighted sum of the cohort-year effects equals the OLS TWFE estimate. The results are presented in Figure 4, where we report the weights used in the computation of the aggregate treatment effects of the ETWFE from Section 2.2.2 in dark color ('ETWFE') along with implicit weights attached by the OLS TWFE estimator to cohort-year cells computed following de Chaisemartin and D'Haultfœuille (2020) in light color ('TWFE').

Figure 4a shows the results aggregated by event time. The aggregation scheme for the target parameter in Section 2.2.2 gives equal weight to every post-treatment observation. As a result, the weights of the ETWFE estimator decrease by event time as the number of cohorts with data for the corresponding number of post-treatment years declines towards the end of the sample. In comparison, the implicit weights of the TWFE are more than two times larger in the first treatment year and then decline much more rapidly. After ten years, the weights of the TWFE are lower than those of the ETWFE estimator. They drop to zero at around 20 years after treatment onset and become slightly negative thereafter. In sum, the TWFE seems to substantially overweight treatment effects shortly after treatment onset, which tend to be small, and to substantially underweight treatment

Figure 4: Weights of OLS ETWFE and TWFE estimator




(a) By event time

(b) By cohort

(c) By event time (1983–1989 cohort)

(d) By event time (1990–1999 cohort)

(e) By event time (2000–2016 cohort)

**Notes**: The figure reports the weights used in the computation of the aggregate treatment effects of the ETWFE from Section 2.2.2 in dark color ('ETWFE') along with the implicit weights attached by the OLS TWFE estimator to cohort-year cells computed following de Chaisemartin and D'Haultfœuille (2020) in light color ('dynamic TWFE'). Panel (a) reports weights aggregated by event time. Panel (b) reports weights aggregated by cohort. Panels (c)–(e) report weights aggregated by event time for the 1983-1989 cohort, the 1990-1999 cohort, and the 2000-2016 cohort, respectively.

effects further from treatment onset, which tend to be large due to maturation effects (Figure 3a).

Figure 4b shows the results aggregated by cohort. From this perspective, the TWFE estimator overweights late-treated cohorts (with small effects) and underweights early-treated cohorts (with large effects). Looking at weights by event time for different cohorts confirms that early (late) cohorts are more strongly underweighted (overweighted)

in all years, which suggests that the cohort effect is not only mechanically driven by compositional differences in terms of event years (see Figure 4c–e).

Overall, around 13% (60 out of 469) of all cohort-year cells have negative weights summing to a total of -.0284. Therefore, while negative weights are present in this setting, they do not seem to play an important role for driving the differences between the TWFE and ETWFE estimates. Rather, in line with the example in Section 2.2.3, the TWFE is biased downward relative to the target parameter of the ETWFE estimate by putting larger weights on treatment effects shortly after treatment onset and on late cohorts, which are associated with small treatment effects.

# 5 Robustness Analysis

To demonstrate the robustness of our main results, we perform a battery of sensitivity experiments. To ease exposition and add structure to the analysis, we split the robustness checks into four groups covering (i) the degree of heterogeneity of the estimator, (ii) potential incidental parameter problems, (iii) DiD methods, and (iv) gravity-related experiments.

## 5.1 Degree of heterogeneity of the estimator

With regard to the degree of heterogeneity of the ETWFE estimator, we either impose restrictions on the treatment effect heterogeneity in the estimation, i.e., the coefficient $\delta_{gs}$ in equation (3), or, alternatively, we also consider more flexible specifications or estimators that allow for more heterogeneity or even provide direct estimates of individual-level heterogeneity. Our findings are reported in Table 4.

**Restrictions on heterogeneity.** First, we impose strong restrictions on the model by allowing the treatment effect to vary only across event time (column (1)) or across cohorts (column (2)). Note that the specification in column (1) is akin to a standard event-study

or dynamic TWFE regression without including leads of the intervention or restricting the event window (cf. Figure 3a). The treatment effect in column (1) is still substantially larger than the static TWFE estimate (0.166 from column (1) in Table 3), yet it is also significantly smaller than the ETWFE baseline estimate, consistent with the results in Figure 3a. The implication is that cohort heterogeneity plays an important role for the RTA estimates, and this is consistent with findings from the gravity literature, e.g., Baier et al. (2019). By contrast, the estimate allowing for only cohort-specific heterogeneity in column (2) is larger than the ETWFE baseline estimate, even though the difference is not statistically significant. In combination, the estimates from columns (1) and (2) highlight the strong treatment effect heterogeneity along the cohort dimension, while treatment effect dynamics seem to be, from this perspective, of second order in this case.

Table 4: Robustness with regard to degree of heterogeneity of the ETWFE estimator and incidental parameter problems

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.247*** | 0.419*** | 0.495*** | 0.433*** | 0.720*** | 0.719 | 0.344 | 0.200*** | 0.417*** |
| | (0.059) | (0.043) | (0.040) | (0.041) | (0.156) | (.) | (.) | (0.067) | (0.090) |
| Estimator | ETWFE | ETWFE | ETWFE | ETWFE | ETWFE | Imputation | Imputation | Jackknife TWFE | Jackknife ETWFE |
| Unit heterogeneity | | Cohort | Coh× RTAID | Cohort | Cohort | Pair | Pair | Cohort | Cohort |
| Time heterogeneity | Year | | 5yr | Year | Year | Year | Year | Year | Year |
| Observations | 105,409 | 105,409 | 89,972 | 89,972 | 105,409 | 104,685 | 104,685 | 105,409 | 105,409 |
| Exporters | 69 | 69 | 67 | 67 | 69 | 69 | 69 | 69 | 69 |
| Importers | 69 | 69 | 67 | 67 | 69 | 69 | 69 | 69 | 69 |
| Years | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 33 | 30 | 820 | 469 | 469 | | | 469 | 469 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes |
| Importer × year FE | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes |
| Cross-border × year FE | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes |
| Year FE | | | | | Yes | Yes | | | |

**Notes**: The table presents PPML regression results using variants of the ETWFE estimator (equation (3)), for which the cohort-time-specific treatment effects were aggregated using equation (7) to obtain an aggregate treatment effect estimate. The dependent variable is exports which vary over the exporter-importer-year dimension. Column (1) imposes complete homogeneity along the cohort dimension and only allows for event-time-specific treatment effects. Column (2) imposes complete homogeneity along the time dimension and only allows for cohort-specific treatment effects. Column (3) allows for more heterogeneity along the cohort dimension by interacting the cohort dummy with an an agreement dummy, while restricting time heterogeneity to 5-year intervals. Column (4) uses the baseline specification, but restricts the sample to be the same as in column (3). Columns (5) and (6) show results using the ETWFE and imputation estimator with only exporter × importer FE and year FE, respectively. Column (7) shows results using the imputation estimator with the same fixed effect structure as the baseline. Columns (8) and (9) show results for the jackknife TWFE and ETWFE using 1,000 draws described in Section 5.2. 'Coefficients' reports the number of estimated coefficients apart from the fixed effects. Standard errors in parentheses are clustered by country pair. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

**Allowing for more heterogeneity.** While – similar to other estimators proposed in the literature (e.g., de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021) – the model in equation (3) only allows identification of (simple) average treatment effects at the cohort-period level, it does not require treatment effects to be homogeneous within cohort-period cells as long as the estimation target is a weighted sum of (average) cohort-period effects as considered in our paper (and the cited papers). However, if interest lies in different estimation targets that require not just average treatment effects of cohort-year cells, but more granular treatment effects, then one needs to allow for "more heterogeneity" ex ante in the model by adding additional interactions. Motivated by this, we also consider an additional specification in column (3), in which we allow for more heterogeneity along the cohort dimension by interacting the cohort dummy with an indicator for denoting individual agreements, while restricting the time heterogeneity to 5-year intervals for computational reasons.[31] The treatment effect of this specification increases relative to the baseline estimate. However, part of the difference is also driven by differences in sample composition – since agreement information is not available for all RTAs in our baseline sample – as suggested by the estimate in column (4), which uses the baseline specification on the sample from column (3).

**Imputation estimator.** An imputation estimator even allows individual treatment effects to be identified (Borusyak et al., 2023). Wooldridge (2023) shows that the imputation approach and the ETWFE estimator are generally not the same for non-linear difference-in-differences, but that they yield numerically equivalent treatment effects when the canonical link function is in the linear exponential family like in the case under consideration. We first set out to confirm this equivalence result by considering the same specification as in Wooldridge (2023), i.e., by including only pair and year fixed effects. As expected, the point estimates of the imputation estimator in column (5) are numerically very close to the one of the ETWFE in column (6) confirming the results obtained

[31] Note that adding a large number of additional coefficients increases the likelihood of the estimation suffering from an incidental parameter problem. In principle, we could have also allowed for pair-specific heterogeneity, but did not do so for computational reasons.

by Wooldridge (2023).[32]

The estimate in column (7) is from an imputation estimator using the richer fixed effect structure from our main specification. In this case, the imputation estimate is slightly smaller than the ETWFE estimate, suggesting that the equivalence breaks down under the more complex fixed effect structure commonly used in the gravity setting. The difference between the ETWFE and the imputation estimate likely stems from the fact that the imputation estimator estimates the fixed effects only using the control group, while the ETWFE estimator uses information on both the control and the treatment group. In case the fixed effects coefficients are different between control and treatment group ex ante or because they are affected by the treatment, it may therefore be more suitable to use the ETWFE estimator.[33] However, most important for current purposes, the imputation estimate is still around twice as large as the associated TWFE estimate. Thus, overall, the additional results using the imputation estimator reinforce our main finding.

In sum, we conclude that the heterogeneity of the ETWFE may be restricted to a certain extent along the time dimension in this setting without appreciable effects on the aggregate treatment effect. This may, of course, not be true for more disaggregated treatment effects, such as cohort-specific treatment effects. Allowing for more heterogeneity by including additional interactions or using an imputation approach slightly changes the point estimate of the aggregate treatment effect, but not the main conclusion that the TWFE estimate is substantially smaller.

[32] While we do not report standard errors for the imputation estimator, they could likely be easily obtained using a bootstrap procedure.

[33] Trade theory suggests that the coefficients on the country-time fixed effects in standard gravity regressions, such as ours, can be very different between control and treatment group both ex ante or because they are affected by the treatment, i.e., due to changes in size and prices/multilateral resistance, which are also a function of trade policy. Therefore, the identifying assumption of the imputation estimator in Borusyak et al. (2023) that "the $X_{it}$ have to be unaffected by the treatment and strictly exogenous to be included in the specification" may not necessarily hold in the three-way gravity setting.

## 5.2 Potential incidental parameter problems (IPPs)

Next, we consider potential IPPs which may arise in non-linear models with fixed effects. This point has attracted significant and ongoing attention in the related trade literature, where the current consensus seems to be that the PPML estimator with two-way fixed effects is asymptotically unbiased even when the time dimension of the panel is fixed (Fernández-Val and Weidner, 2016), while the three-way PPML estimator may be asymptotically biased for small $T$ due to the IPP (Weidner and Zylkin, 2021). Furthermore, estimates of cluster-robust sandwich-type standard errors may be downward biased in both two-way and three-way gravity settings (Jochmans, 2017; Pfaffermayr, 2021; Weidner and Zylkin, 2021).

**Monte Carlo simulation.** To study the potential IPP of the ETWFE and TWFE, we implement a Monte Carlo simulation closely following Weidner and Zylkin (2021) to study the potential bias and coverage properties of the TWFE and ETWFE estimator in our setting.[34] The results are computed using 1,000 repetitions and displayed in Table 5. For the TWFE estimator, we find that the bias on the RTA coefficient is zero in this setting. For the ETWFE estimator, the bias is very small ($-0.003$) relative to the coefficient of interest (0.5). With regard to the coverage probability, the standard error estimates of ETWFE seem to be downward biased (0.922), while the standard error estimate of the TWFE estimator is only slightly below 0.95 (0.941), i.e., the value expected for an unbiased estimator. In sum, we conclude that, given the relatively large time dimension considered in our setting, IPP might be less of a problem for the coefficient estimate of the ETWFE, while the associated standard error is potentially downward biased.[35]

[34]For the simulation analysis, we assume the same data generating process as Weidner and Zylkin (2021), but we add cross-border × year fixed effects drawn from a normal distribution with mean zero and a variance of 1/16 (as for the remaining fixed effects). We focus on a "log-homoskestic" variance of the error term (DGP III in Weidner and Zylkin (2021)) studied in Santos Silva and Tenreyro (2006) and a sample of $N = 69$ countries and $T = 34$ years like in our baseline sample. We assume that from $t = 3$ onwards 30 RTAs (drawn at random without replacement) are signed every year, which results in a similar ratio between the treatment and the never-treated group as in our baseline sample. Accordingly, the independent variable $x_{ijt}$ is determined and $\beta$ is set to 0.5, i.e., in the simulation analysis, we assume treatment effect homogeneity across cohorts and time.

[35]For our baseline, we estimate 469 coefficients with 19,642 post-treatment observations in the treatment

Table 5: Monte Carlo simulation

| Estimator | Average bias | Coverage probability |
|---|---|---|
| PPML TWFE | 0.000 | 0.941 |
| PPML ETWFE | -0.003 | 0.922 |
| PPML TWFE jackknife | 0.000 | 0.950 |
| PPML ETWFE jackknife | -0.000 | 0.938 |

**Notes**: The table presents the results of the Monte Carlo simulation described in Section 5.2 using 1,000 repetitions. Average bias refers to the mean of the difference between $\hat{\beta}$ and $\beta$. Coverage probability refers to the probability that $\beta = 0.5$ is covered in the 95% confidence interval for $\hat{\beta}$, which should be 0.95 for an unbiased estimator. Jackknife refers to a split-sample jackknife estimate (Weidner and Zylkin, 2021).

As a potential remedy, we consider a version of the (split-panel) jackknife studied in Dhaene and Jochmans (2015), Pfaffermayr (2021), and Weidner and Zylkin (2021) for bias correction. The resulting jackknife TWFE estimator shows zero bias and a perfect coverage probability of 0.95. Similarly, the jackknife ETWFE estimator also has zero bias and a coverage probability of slightly under, but close to 0.95 (0.938). We conclude that the jackknife might help with any remaining bias of the ETWFE estimator and also exhibits approximately correct coverage probabilities.

**Jackknife bias correction.** Motivated by the simulation results, we also apply the TWFE and ETWFE jackknife estimator to our baseline data set (column (8) and (9) in Table 4). For the jackknife TWFE, we obtain an RTA coefficient of 0.200 (baseline: 0.166) with a standard error of 0.067 (baseline: 0.050),[36] while for the jackknife ETWFE, we obtain an RTA estimate of 0.417 (baseline 0.381) with a standard error of 0.090 (baseline: 0.041). This suggests that – possibly due to differences in the data generating process relative to the case considered in the simulation analysis – the coefficient estimates of TWFE and ETWFE in our baseline might both be slightly downward biased. In line with the simulation analysis, the standard error estimate of the baseline ETWFE seems to be downward biased more than the standard error estimate of the baseline TWFE. The resulting jackknife ETWFE standard error estimate is now larger than the

group (105,409 observations in total), i.e., around 42 (225) observations per coefficient.

[36] Using the analytical bias correction of Weidner and Zylkin (2021) also indicates that the TWFE coefficient and standard error estimates are downward biased (see also Figure 1).

jackknife TWFE standard error estimate, consistent with the intuition that a more flexible estimator comes at the cost of precision.

Importantly, however, the overall conclusion that the ETWFE estimator yields an RTA coefficient twice as large as the TWFE estimator remains unchanged. A formal statistical test (Z-test) that takes the covariance between the estimates into account yields a p-value of 0.013, i.e., the null hypothesis of equality of coefficients is rejected at the 5% level. Based on this analysis, we recommend computing jackknife coefficient and standard error estimates for the ETWFE in the three-way gravity setting at least as a robustness check.

## 5.3 DiD-related experiments

This subsection offers results from five sets of experiments related to the implementation and robustness of our methods. First, we provide additional robustness checks with regard to the degree of heterogeneity of the time dimension of the ETWFE estimator. Second, we use alternative control groups. Third, we consider an extension of the ETWFE estimation approach in which the cohort-time-specific treatment effects are allowed to vary by time-constant covariates. Fourth, we test the robustness of our results to the choice of the treatment onset. Lastly, we experiment with alternative weighting schemes.

**Degree of heterogeneity.** With regard to the heterogeneity of the ETWFE estimator, we experiment with additional mild restrictions along the time dimension. Our findings are reported in Table 6. First, in column (1), we use a specification with binned endpoints (e.g., Schmidheiny and Siegloch, 2020), i.e., in which we restrict the treatment effect to remain constant after 10 years in line with the results in Egger et al. (2022). Second, in columns (2) and (3), we allow treatment effects to only change every 2 years or every 5 years in the spirit of the practice of estimating gravity equations with interval or averaged data, e.g., Baier and Bergstrand (2007) and Olivero and Yotov (2012), while employing consecutive year data. All these three adjustments leave the treatment effect largely unaffected, while strongly reducing the number of parameters to be estimated.

Table 6: Robustness with regard to different DiD-specific assumptions

| | Baseline | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.381*** | 0.393*** | 0.382*** | 0.383*** | 0.322*** | 0.326*** | 0.278*** | 0.436*** | 0.539*** | 0.535*** | 0.324*** | 0.459*** | 0.440*** |
| | (0.041) | (0.041) | (0.041) | (0.041) | (0.036) | (0.037) | (0.053) | (0.045) | (0.052) | (0.054) | (0.039) | (0.051) | (0.047) |
| *Heterogeneity* | | | | | | | | | | | | | |
| Unit heterogeneity | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort | Cohort |
| Time heterogeneity | Year | Year | 2yr | 5yr | Year | Year | Year | Year | Year | Year | Year | Year | Year |
| Binning | | 10yr+ | | | | | | | | | | | |
| *Control group* | | | | | | | | | | | | | |
| Not-yet treated | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Never treated | Yes | Yes | Yes | Yes | Yes | Yes | | Yes | Yes | Yes | Yes | Yes | Yes |
| *Omit anticipation periods* | | | | | | | | Yes | | | | | |
| *Covariate interactions* | | | | | | | | | | | | | |
| ln Distance | | | | | | | | | Yes | Yes | | | |
| Contiguity | | | | | | | | | | Yes | | | |
| Colony | | | | | | | | | | Yes | | | |
| Language | | | | | | | | | | Yes | | | |
| *Weights* | Obs | Obs | Obs | Obs | Obs | Obs | Obs | Obs | Obs | Obs | Cohort | Year | Cohort × year |
| Observations | 105,409 | 105,409 | 105,409 | 105,409 | 105,409 | 89,280 | 34,414 | 102,022 | 100,681 | 100,681 | 105,409 | 105,409 | 105,409 |
| Exporters | 69 | 69 | 69 | 69 | 69 | 69 | 66 | 69 | 68 | 68 | 69 | 69 | 69 |
| Importers | 69 | 69 | 69 | 69 | 69 | 69 | 66 | 69 | 68 | 68 | 69 | 69 | 69 |
| Years | 34 | 34 | 34 | 34 | 34 | 34 | 33 | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 469 | 255 | 242 | 107 | 984 | 442 | 439 | 469 | 971 | 2,477 | 469 | 469 | 469 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Importer × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cross-border × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

**Notes**: The table presents PPML regression results using variants of the ETWFE estimator (equation (3)), for which the cohort-time-specific treatment effects were aggregated using equation (7) to obtain an aggregate treatment effect estimate. The dependent variable is exports which vary over the exporter-importer-year dimension. The 'Baseline' estimate is based on equation (3). Column (1) restricts the cohort-time-specific treatment effects to remain unchanged after ten or more years after treatment onset. Columns (2) and (3) restrict the cohort-time-specific treatment effects to change only every two or every five years. Columns (4) and (5) limit the control group to include never-treated country pairs by saturating all pre-treatment observations of not-yet-treated country pairs with cohort-year-specific fixed effects (column (4)) or by dropping the not-yet-treated observations (apart from the two necessary pre-treatment observations to identify the corresponding treatment effect (Sun and Abraham, 2021; Borusyak et al., 2023) from the sample, respectively (column (5)). Column (6) limits the control group to include not-yet-treated country pairs by dropping all never-treated observations from the sample. Column (7) omits the three years before RTAs' entry into force in treated country pairs following Wooldridge (2023). Columns (8) and (9) include interactions between ln Distance and cohort-time-specific treatment effects (column (8)) and ln Distance, Contiguity, Colony, Language, and cohort-time-specific treatment effects (column (9))) thereby relaxing the parallel trend assumption (Callaway and Sant'Anna, 2021; Wooldridge, 2023). Columns (10)-(12) report results for alternative weighting schemes that differ from the approach in all other specifications that give every post-treatment observation ('Obs') the same weight. Instead, the robustness checks give every cohort (column (10)), every event year (column (11)), or every cohort-year (column (12)) the same weight. 'Coefficients' reports the number of estimated coefficients apart from the fixed effects. Standard errors in parentheses are clustered by country pair. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

**Alternative control groups.** Next, we experiment with alternative control groups. Our findings are reported in Table 6. For the baseline estimate, the estimating sample consists of the two groups: never-treated country pairs and non-yet-treated country pairs. Never-treated country pairs are those in which no RTA entered into force in our sample, i.e., between 1980 and 2016. Not-yet-treated country pairs are those with no RTA onset until the year of the comparison, but did so in later years of the sample.

First, we only use the never-treated group as a control group by saturating all pre-treatment observations of not-yet-treated country pairs with cohort-year-specific fixed effects (column (4)) or by dropping the not-yet-treated observations (apart from the two necessary pre-treatment observations to identify the corresponding treatment effect (Sun and Abraham, 2021; Borusyak et al., 2023) from the sample (column (5)). Both estimates are slightly smaller than the baseline estimate. However, the difference is not statistically significant.

Second, we only use the not-yet-treated group as a control group by dropping all never-treated observations from the sample (column (6)). Note that this comes at a loss of efficiency due to the smaller number of observations and does not allow identification of treatment effects for the last treatment cohort. The resulting RTA estimate is significantly smaller than the ETWFE baseline estimate. On the one hand, the not-yet-treated group might be a better control group than the never-treated group in the sense that it is more similar to the treatment group since the associated country pairs also sign RTAs in later years. On the other hand, the never-treated group is by definition unaffected by potential anticipation effects. In sum, the baseline RTA effect appears mainly driven by comparisons with never-treated country pairs and the estimate might be somewhat smaller when limiting the control group to not-yet-treated country pairs.

**Treatment onset.** We conclude the analysis in this section with a robustness test for anticipation effects. Instead of assuming an 'onset' of RTAs three years before their entry into force (see Section 3), we omit these time periods in treated country pairs following Wooldridge (2023). Unsurprisingly given the time profile of the RTA effects, the resulting

treatment effect estimate in column (7) of Table 6 is slightly larger than our baseline estimate. This is due to the fact that the first three initial years are omitted which are associated with low cohort-time-specific treatment effects capturing short-term rather than long-term effects of RTAs. We conclude that our RTA estimates are robust to our definition of RTA onset.

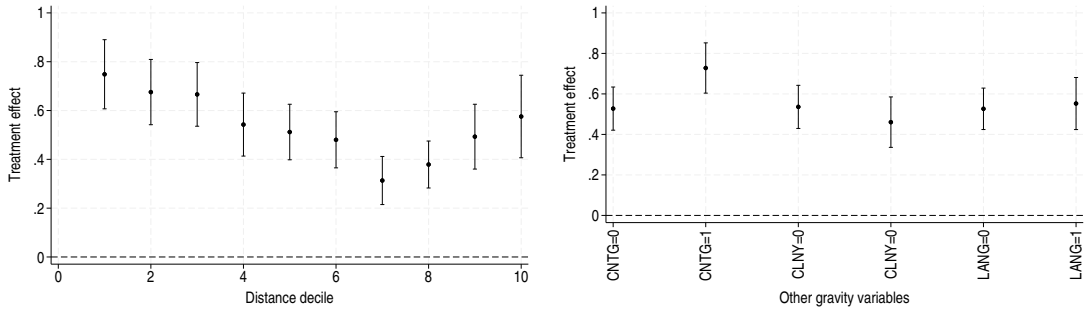
**Time-constant covariates.** Next, we consider an extension of the ETWFE estimation approach in which the cohort-time-specific treatment effects are allowed to vary by time-constant covariates (Wooldridge, 2023). As discussed in Wooldridge (2021), this is a parametric version of the regression adjustment approach by Heckman et al. (1997) for panel data. This specification relaxes the parallel trend assumption, which now only needs to hold conditional on covariates, thereby rendering it more plausible. This is similar in spirit to the approach by Callaway and Sant'Anna (2021) who consider settings when the parallel trends assumption only holds after conditioning on observed covariates by using outcome regression, inverse probability weighting, and doubly-robust estimands. As time-constant covariates, we consider standard bilateral gravity variables.

Column (8) reports an estimate using the distance between country pairs and column (9) the distance in combination with contiguity, language, and past colonial relations. Both estimates are very similar (0.539 vs. 0.535) and substantially larger than the baseline RTA estimate. This provides suggestive evidence that making treatment and control groups more comparable, in particular, with regard to distance and thereby relaxing the parallel trends assumptions may result in significantly larger treatment effects.

To better understand the larger RTA estimate in this specification, we first computed treatment effects for different values of the covariates (Figure 5). The treatment effect decreases by distance up to the seventh decile and then increases again slightly. This is, in principle, in line with Baier et al. (2018), who find a negative coefficient on the interaction between distance and RTAs in a TWFE specification. The (negative) impact of distance on the RTA effect of trade could be related to variable transport costs, but should be interpreted with caution as distance could also be strongly correlated with other

explanatory variables not included in the specification. We also computed the treatment effects for different values of contiguity, colony, and language. We find that the treatment effect for contiguity is slightly larger (albeit not significantly so) in line with the result on distance, while the treatment effects split by colony and language turn out to be very similar.

Figure 5: Treatment effect by covariates

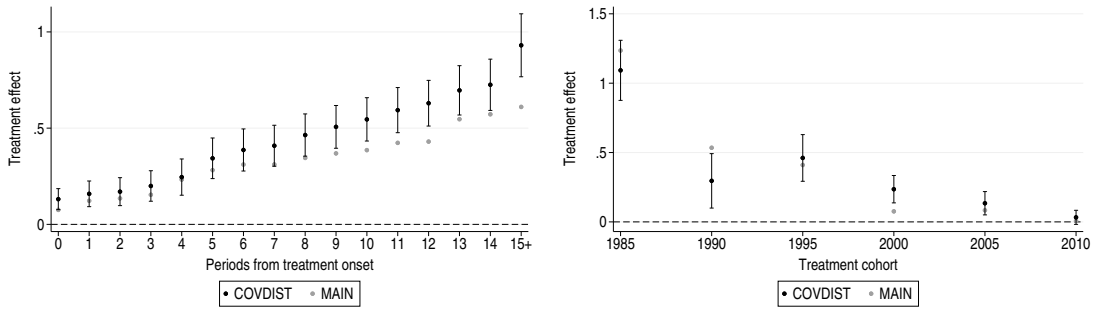



(a) By distance (b) By other gravity variables

**Notes**: The figure reports treatment effects aggregated using equation (7) for different levels of the covariates interacted with the cohort-year dummies for the specification in column (8) (panel (a)) and column (9) (panel (b)) of Table 6. Panel (a) reports treatment effects for different deciles of the variable "Distance". Panel (b) reports treatment effects for the indicator variables contiguity (CNTG), colony (CLNY), and common language (LANG). 95% confidence intervals are shown using standard errors clustered by country pair.

Second, we compute treatment effects by event time and by cohort group and compare them to the results from our baseline specification (Figure 6). Regarding the results by event time, we find that controlling for distance leads to larger treatment effects, in particular, in periods far away from treatment onset and makes the RTA effects on trade longer lasting. Interestingly, regarding the results by cohort, we find that the effect in early-treated cohorts becomes slightly smaller and the effect in late-treated cohorts becomes slightly larger. This leads to a reduction in the heterogeneity of the RTA effect across cohorts that we find in the baseline, suggesting that heterogeneity by covariates might be one factor driving the differences in RTA effectiveness across cohorts (see also Section 4.3 for a more detailed discussion).

Figure 6: Event-time-specific and cohort-specific treatment effects



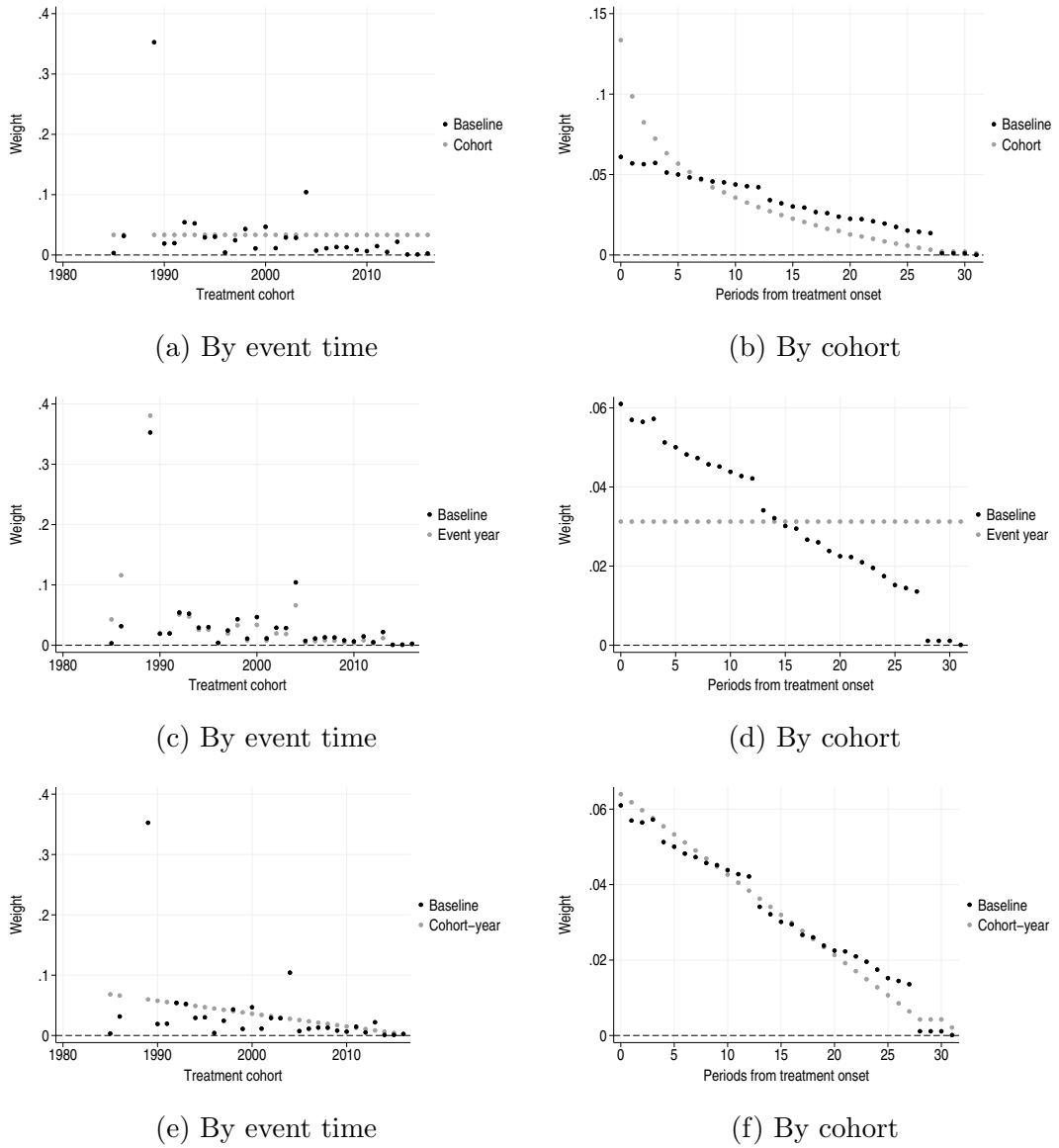

(a) By event time (b) By cohort (5-year bins)

**Notes**: The figure reports different aggregations of the cohort-year-specific treatment effects from PPML estimation of equation (3). Panel (a) reports event-time-specific treatment effects from equation (3) aggregated using equation (9) from the covariate-augmented specification of column (8) in Table 6 in dark color ('COVDIST') along with the baseline specification of column (2) in Table 3 in light color ('MAIN'). Panel (b) reports cohort-specific treatment effects from equation (3) aggregated using equation (8) from the covariate-augmented specification of column (8) in Table 6 in dark color ('COVDIST') along with the baseline specification of column (2) in Table 3 in light color ('MAIN'). 95% confidence intervals are shown using standard errors clustered by country pair.

In sum, it is reassuring that the results regarding covariate heterogeneity are in line with the previous literature and that relaxing the identifying assumptions yields results that tend to be larger than our baseline specification, reinforcing our key result that the effects of RTA are larger than commonly thought.

**Alternative weighting schemes.** In our next set of DiD-related experiments, we use alternative weighting schemes. For our target parameter, we give every post-treatment observation the same weight. To reduce the potential impact of sample composition and limit the influence of individual agreements on the aggregate result, as a robustness check, we give every cohort, every event year, or every cohort-year the same weight. The resulting weights by event time and by cohort are displayed in Figure 7 along with the weights of our baseline ETWFE estimate, and the corresponding aggregate treatment effects are reported in columns (10)-(12) of Table 6.

Figure 7: Weights of the baseline PPML ETWFE and alternative weighting schemes




(a) By event time

(b) By cohort

(c) By event time

(d) By cohort

(e) By event time

(f) By cohort

**Notes**: The figure reports the weights used in the computation of the aggregate treatment effects of the ETWFE from Section 2.2.2 in dark color ('ETWFE') along alternative weighting schemes in light color. In this regard, Panel (a)–(b), (c)–(d), and (e)–(f) report weights of a weighting scheme, which gives every cohort, every event year, or every cohort-year the same weight, respectively.

This analysis reveals that the aggregate treatment effects are slightly smaller (0.324) for the weighting scheme that gives every cohort the same weight since cohorts with large average effects have a large number of observations in our sample. By contrast, the aggregate treatment effect is larger (0.459 and 0.440) for the weighting schemes that give every event year or every cohort-year the same weight. In the first case, this results from the fact that the large treatment effects further away from treatment onset are given a

larger weight than in the baseline. In the second case, this stems from larger weights for earlier cohorts (that by definition have more distinct cohort-year pairs), which, on average, show larger treatment effects.

## 5.4 Gravity-related experiments

In this subsection, we explore whether the ETWFE estimator is more (or less) sensitive to the standard set of robustness checks from the gravity literature. To this end, we perform nine robustness experiments and, in each of them, we rely on our main econometric specification while only changing one feature of the estimating model or the estimating sample at a time. Similar to the main analysis, in each of the new experiments, we obtain and report two sets of TWFE and ETWFE estimates. Then, we compare them against each other and also against the corresponding benchmark results from Table 3. The main results from our gravity-related experiments are reported in Table 7.

**OLS estimator.** We start by reproducing our main results from columns (1) and (2) of Table 3 with the OLS estimator. The motivation for the OLS specification is twofold. Even though PPML has established itself as the leading gravity estimator (e.g., Santos Silva and Tenreyro, 2006, 2021), there are still many researchers who estimate gravity with OLS or, at least, report OLS estimates as a robustness check. In addition, as discussed earlier, most of the recent heterogeneity-robust staggered DiD methods are implemented in linear settings (e.g., Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; Borusyak et al., 2023; de Chaisemartin and D'Haultfoeuille, 2022). Thus, a comparison between the RTA estimates obtained with the OLS and PPML estimators could be beneficial from that perspective too.

Table 7: Robustness with regard to different gravity specifications

| | (1) TWFE | (2) ETWFE | (3) TWFE | (4) ETWFE | (5) TWFE | (6) ETWFE | (7) TWFE | (8) ETWFE | (9) TWFE | (10) ETWFE | (11) TWFE | (12) ETWFE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.172*** | 0.347*** | 0.143*** | 0.392*** | 0.022 | 0.186*** | 0.158*** | 0.378*** | 0.481*** | 1.091*** | 0.174* | 0.279*** |
| | (0.037) | (0.051) | (0.051) | (0.044) | (0.038) | (0.034) | (0.048) | (0.039) | (0.089) | (0.174) | (0.091) | (0.058) |
| $WTO_{ij,t}$ | | | | | | | 0.331*** | | | | | |
| | | | | | | | (0.061) | | | | | |
| $DIST_{ij}$ | | | | | | | | | -0.312*** | | | |
| | | | | | | | | | (0.077) | | | |
| $CNTG_{ij}$ | | | | | | | | | 1.019*** | | | |
| | | | | | | | | | (0.196) | | | |
| $LANG_{ij}$ | | | | | | | | | 0.423*** | | | |
| | | | | | | | | | (0.101) | | | |
| $CLNY_{ij}$ | | | | | | | | | 0.220 | | | |
| | | | | | | | | | (0.134) | | | |
| $GDP_{i,t}$ | | | | | | | | | | | 1.324*** | |
| | | | | | | | | | | | (0.192) | |
| $GDP_{j,t}$ | | | | | | | | | | | 0.885*** | |
| | | | | | | | | | | | (0.196) | |
| $REM_{j,t}$ | | | | | | | | | | | -0.328 | |
| | | | | | | | | | | | (0.555) | |
| $REM_{i,t}$ | | | | | | | | | | | -0.319 | |
| | | | | | | | | | | | (0.570) | |
| Estimator | OLS | OLS | PPML | PPML | PPML | PPML | PPML | PPML | PPML | PPML | PPML | PPML |
| Domestic trade | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes | Yes | Yes | Yes |
| 5-yr interval | | | Yes | Yes | | | | | | | | |
| Observations | 104,818 | 104,818 | 20,854 | 20,854 | 103,530 | 103,530 | 105,395 | 105,395 | 100,682 | 100,682 | 99,953 | 99,953 |
| Exporters | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 68 | 68 | 67 | 67 |
| Importers | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 68 | 68 | 67 | 67 |
| Years | 34 | 34 | 7 | 7 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 1 | 469 | 1 | 93 | 1 | 469 | 2 | 470 | 5 | 473 | 5 | 473 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | | | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | | |
| Importer × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | | |
| Cross-border × year FE | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes | Yes | Yes | Yes |

**Notes**: The table presents regression results using the TWFE estimator (equation (1)) and the ETWFE estimator (equation (3)), for which the cohort-time-specific treatment effects were aggregated using equation (7) to obtain an aggregate treatment effect estimate. The dependent variable is exports which vary over the exporter-importer-year dimension. 'Estimator' indicates whether results were obtained using the OLS or the PPML estimator. 'Domestic trade' indicates whether domestic trade flows were included in the sample or not. '5-yr interval' indicates whether 5-year interval data was used in the estimation. 'Covariate controls' reports the covariates that were added as controls in the estimation. Exporter and importer remoteness are atheoretical proxies for the structural multilateral resistances computed as exporter or importer GDP-weighted bilateral distances. 'Coefficients' reports the number of estimated coefficients apart from the fixed effects. Standard errors in parentheses are clustered by country pair. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

We draw two main conclusions based on the OLS estimates from columns (1) and (2) of Table 7. First, the TWFE estimate in column (1) of Table 7 is very close to the corresponding PPML result from column (1) of Table 3. Second, the ETWFE OLS estimate from column (2) of Table 7 is a bit smaller than the corresponding PPML result from Table 3, yet it is still more than twice as large as the TWFE OLS estimate from column (1) of Table 7. Thus, our main conclusions that the TWFE gravity estimates may be biased downward and that the heterogeneity-robust staggered DiD methods deliver estimates that are more consistent with policy expectations are confirmed with the OLS estimator.

**Interval data.** Motivated by the tradition in the trade literature of estimating the gravity equation with interval (instead of consecutive-year) data (e.g., Cheng and Wall, 2005; Baier and Bergstrand, 2007; Olivero and Yotov, 2012), in our next experiment we use 5-year interval data. Our findings are reported in columns (3) and (4) of Table 7. Even though the new TWFE estimate is a bit smaller than our main estimate from column (1) of Table 3, and the new ETWFE estimate is a bit larger than the corresponding estimate from column (1) of Table 3, we view the interval-data results as comparable to our main findings, thus confirming the bias in the TWFE estimates.

Despite the similar results that we obtain with the consecutive-year and the interval data, and as also argued in Egger et al. (2022), we recommend the use of consecutive-year data because the interval estimates may miss some of the adjustments in response to the formation of RTAs. In addition, we believe that using all possible years in the data is especially beneficial in staggered DiD settings not only from an estimation efficiency perspective, but also because this would enable researchers to more precisely estimate the underlying cohort-time-specific treatment effects that provide additional information. Thus, a further implication of our analysis is that the use of the ETWFE estimator provides an additional argument against using interval data in gravity regressions.

**Domestic trade flows.** As discussed in Yotov (2022), there may be significant benefits

of estimating the gravity model with domestic (in addition to international) trade flows. Nevertheless, most trade gravity regressions are estimated with data on international trade flows only.[37] Therefore, in our next experiment, we only use data on international trade flows. The results are reported in columns (5) and (6) of Table 7. Two findings stand out. First, both the TWFE and the ETWFE estimates from Table 7 are significantly smaller than their counterparts from Table 3. In fact the TWFE estimate is no longer statistically significant. This result is consistent with estimates from the RTA literature (e.g., Dai et al., 2014; Baier et al., 2019; Larch and Yotov, 2023), and the intuition for the larger RTA estimates from the sample with domestic trade flows is that the estimates of trade agreements that are based on international trade flows only may be biased downward because they cannot capture diversion from domestic sales.

Second, and more important for our purposes, we see that, even though the ETWFE estimate in column (6) is half the size of the corresponding result from Table 3, it is still significantly larger than the TWFE estimate from column (5) of Table 7, thus confirming our main result about the potential bias in the TWFE gravity estimates. We also note that, unlike the TWFE estimate, the ETWFE estimate is statistically significant. A potential implication of this analysis for gravity estimations is that the ETWFE estimates may not be as sensitive as the TWFE estimates to the addition of domestic trade flows to the estimating sample.

**GATT/WTO membership.** In our next experiment, we control for GATT/WTO membership.[38] The motivation for this specification is that omitting the impact of WTO may bias the RTA estimates upwards, e.g., because the latter may capture common globalization effects that should not be attributed to the RTAs. The results are reported in columns (7) and (8) of Table 7 and support our main conclusions. Specifically, we see

[37] Traditionally, this is due to lack of data on domestic trade flows. Data on domestic sales have recently become more widely available and more reliable. Therefore, we see more estimations in gravity analysis that are performed on samples that combine international and domestic sales.

[38] Note that in this and the following specifications in this subsection, we simply add covariates as controls, i.e., we do not interact them with the cohort-time-specific treatment effects like we did in columns (7) and (8) of Table 6. In doing so, we slightly diverge from the approach described in Wooldridge (2023), but adopt the standard that is used in the gravity literature.

that, even though both the TWFE and the ETWFE estimates are a bit smaller than our main results from Table 3, neither of the new estimates are affected significantly by the introduction of the control variable for GATT/WTO membership. Importantly, the difference between the TWFE or the ETWFE estimates remains large and in favor of the latter.

**Standard gravity variables.** The results in columns (9) and (10) of Table 7 are obtained after replacing the pair fixed effects from our main specification with the set of 'standard' gravity variables, including the log of bilateral distance, and dummy variables for sharing common borders, common language, and colonial ties. The resulting TWFE and ETWFE estimates are significantly larger than the corresponding main estimates from columns (1) and (2) of Table 3. More important for our purposes, the gap between the new ETWFE and TWFE estimates is similar to that from our main analysis (i.e., the ETWFE estimate is more than twice larger than the TWFE estimate), thus, once again, confirming our main conclusions.

**Exporter-time and importer-time fixed effects.** Next, we estimate a specification that does not include exporter-time and importer-time fixed effects. In principle, we would not recommend this specification from the perspective of the structural gravity literature, because it does not control properly for the theoretical multilateral resistances, which is considered a 'gold medal mistake' in gravity estimations (Baldwin and Taglioni, 2006). Nevertheless, we still perform this analysis for two reasons. First, because, depending on the key covariate of interest, it may not be possible to include the exporter-time and importer-time fixed effects. Second, because we want to check whether the ETWFE estimates respond differently than the TWFE estimates to the omission of the exporter-time and importer-time fixed effects.

The corresponding results appear in columns (11) and (12) of Table 7, where, instead of the exporter-time and importer-time fixed effects, we added as control variables the GDPs of the exporter and of the importer, as proxies for country size, and we constructed

atheoretical proxies for the structural multilateral resistances as GDP-weighted bilateral distances. The TWFE estimate is now only significant at the 10% level and again smaller in magnitude, while the ETWFE estimate is still large and statistically significant. Thus, based on this analysis, our main conclusion that we draw based on these results is that the ETWFE estimates seem to be more robust to omitting certain exporter-time and importer-time characteristics.

**Zero trade flows.** In addition to the main advantage of the PPML estimator, which is to account for potential heteroskedasticity of the trade flows data, the multiplicative form of PPML is very convenient for handling zero trade flows. In our next experiment, we investigate the importance of the presence of zero trade flows for our main findings. To this end, in Table 8 (ETWFE) and Table 9 (TWFE), we reproduce our main results (i.e., of each of the three samples from Table 3) with four alternative specifications. Specifically, columns 'PPML0' report estimates that are obtained after we replaced all missing values with zeros, thus inflating the number of zeros in the sample. For comparison, columns 'PPML' report our main estimates. PPML estimates that are obtained with positive values only are reported in columns 'PPML+'. Lastly, we also provide OLS estimates (in columns 'OLS').

Table 8: Additional results on the difference between PPML and OLS estimates (ETWFE)

| | (1) PPML0 | (2) PPML | (3) PPML+ | (4) OLS | (5) PPML0 | (6) PPML | (7) PPML+ | (8) OLS | (9) PPML0 | (10) PPML | (11) PPML+ | (12) OLS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.362*** | 0.381*** | 0.383*** | 0.347*** | 0.312*** | 0.327*** | 0.328*** | 0.242*** | 0.356*** | 0.293*** | 0.299*** | 0.213*** |
| | (0.039) | (0.041) | (0.041) | (0.051) | (0.038) | (0.040) | (0.040) | (0.043) | (0.040) | (0.039) | (0.038) | (0.032) |
| Sample | Baseline | Baseline | Baseline | Baseline | Medium | Medium | Medium | Medium | Large | Large | Large | Large |
| Observations | 111,625 | 105,409 | 104,818 | 104,818 | 185,125 | 175,796 | 172,645 | 172,645 | 617,312 | 591,092 | 502,370 | 502,370 |
| Exporters | 69 | 69 | 69 | 69 | 91 | 91 | 91 | 91 | 225 | 225 | 225 | 225 |
| Importers | 69 | 69 | 69 | 69 | 91 | 91 | 91 | 91 | 225 | 225 | 225 | 225 |
| Years | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 660 | 469 | 469 | 469 | 660 | 469 | 469 | 469 | 660 | 528 | 528 | 528 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Importer × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cross-border × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

**Notes**: The table presents regression results using the ETWFE estimator (equation (3)), for which the cohort-time-specific treatment effects were aggregated using equation (7) to obtain an aggregate treatment effect estimate. 'PPML0' denotes specifications estimated using PPML, in which the sample was augmented by replacing all missing values in trade flows with zeros, thus inflating the number of zeros in the sample. 'PPML' denotes specifications estimated using PPML. 'PPML+' denotes specifications estimated using PPML, in which the sample was limited to positive trade flows. 'OLS' denotes specifications estimated using OLS. The dependent variable is exports which vary over the exporter-importer-year dimension. The 'Baseline' sample contains 69 countries, accounting for 98% of world exports. The 'Medium' sample contains 91 countries, accounting for 99% of world exports. The 'Large' sample contains the full set of countries from the structural gravity dataset. 'Coefficients' reports the number of estimated coefficients apart from the fixed effects. Standard errors in parentheses are clustered by country pair. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

Table 9: Additional results on the difference between PPML and OLS estimates (TWFE)

| | (1) PPML0 | (2) PPML | (3) PPML+ | (4) OLS | (5) PPML0 | (6) PPML | (7) PPML+ | (8) OLS | (9) PPML0 | (10) PPML | (11) PPML+ | (12) OLS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.160*** | 0.166*** | 0.166*** | 0.172*** | 0.161*** | 0.167*** | 0.167*** | 0.112*** | 0.160*** | 0.165*** | 0.165*** | 0.117*** |
| | (0.050) | (0.050) | (0.050) | (0.037) | (0.048) | (0.048) | (0.048) | (0.031) | (0.047) | (0.047) | (0.047) | (0.023) |
| Sample | Baseline | Baseline | Baseline | Baseline | Medium | Medium | Medium | Medium | Large | Large | Large | Large |
| Observations | 111,625 | 105,409 | 104,818 | 104,818 | 185,125 | 175,796 | 172,645 | 172,645 | 617,312 | 591,092 | 502,370 | 502,370 |
| Exporters | 69 | 69 | 69 | 69 | 91 | 91 | 91 | 91 | 225 | 225 | 225 | 225 |
| Importers | 69 | 69 | 69 | 69 | 91 | 91 | 91 | 91 | 225 | 225 | 225 | 225 |
| Years | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Importer × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cross-border × year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

**Notes**: The table presents regression results using the TWFE estimator (equation (1)). 'PPML0' denotes specifications estimated using PPML, in which the sample was augmented by replacing all missing values in trade flows with zeros, thus inflating the number of zeros in the sample. 'PPML' denotes specifications estimated using PPML. 'PPML+' denotes specifications estimated using PPML, in which the sample was limited to positive trade flows. 'OLS' denotes specifications estimated using OLS. The dependent variable is exports which vary over the exporter-importer-year dimension. The 'Baseline' sample contains 69 countries, accounting for 98% of world exports. The 'Medium' sample contains 91 countries, accounting for 99% of world exports. The 'Large' sample contains the full set of countries from the structural gravity dataset. 'Coefficients' reports the number of estimated coefficients apart from the fixed effects. Standard errors in parentheses are clustered by country pair. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

We see from Tables 8 and 9 that the estimates are a bit different across the different samples and specifications. However, the main conclusion is that the influence of the zeros is not too large. Specifically, for ETWFE, the coefficient for 'PPML0' is slightly smaller for the SMALL and MEDIUM data set and larger for the LARGE data set. For TWFE, there are no apparent and large differences. Thus, our estimates from this experiment reinforce the now standard result in the trade literature that the zeros do not matter too much. Two possible explanations for the small influence of the zeros in gravity estimations are that (i) PPML weights larger observations more, i.e., in effect it discounts the zeros, and (ii) the rich structure of fixed effects in our model (and, in fact, in most of standard gravity regressions from the existing literature) renders most of the zero trade flows absolutely irrelevant for gravity estimations.[39]

**Alternative clustering.** In our next experiment, we investigate the robustness of our results to alternative clusterings of the standard errors. Our results appear in Table 10. The results in column 'Baseline' are clustered by country pair. The results in column (1) are clustered by exporter and importer. The standard errors become slightly larger, but the significance remains unchanged. In column (2), the standard errors are clustered by exporter-year and importer-year, and they become smaller. Lastly, in column (3), the standard errors are clustered by exporter, importer, and year. The standard errors become slightly larger, but the significance of the coefficient estimate remains unchanged. In sum, while alternative clustering seems to matter for the magnitude of the standard errors, the changes are not large and our main results and conclusions remain valid.

**Deep trade agreements.** Larch and Yotov (2023) show that the impact of RTAs may vary by type of agreement. Moreover, Hofmann et al. (2019) and Mattoo et al. (2020) demonstrate that RTAs have become 'deeper' over time, in the sense that more recent

[39]For example, if a country does not produce a product at all, this is accounted for by the exporter-time fixed effects, or if two countries never trade with each other, then this is accounted for by the country-pair fixed effects. Thus, the only relevant zeros in our setting are those where we observe action on the extensive margin of trade, i.e., if trade switches from zero to positive or vice versa. However, there are relatively few such instances with aggregated data.

Table 10: Robustness with regard to clustering of standard errors and additional results on depth of agreements

| | Baseline | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|---|
| $RTA_{ij,t}$ | 0.381*** | 0.381*** | 0.381*** | 0.381*** | | |
| | (0.041) | (0.057) | (0.024) | (0.055) | | |
| $DEPTH_{ij,t} < P50$ | | | | | 0.030 | 0.279*** |
| | | | | | (0.081) | (0.073) |
| $DEPTH_{ij,t} >= P50$ | | | | | 0.181*** | 0.605*** |
| | | | | | (0.053) | (0.042) |
| Estimator | ETWFE | ETWFE | ETWFE | ETWFE | TWFE | ETWFE |
| Observations | 105,409 | 105,409 | 105,409 | 105,409 | 90,369 | 90,369 |
| Exporters | 69 | 69 | 69 | 69 | 67 | 67 |
| Importers | 69 | 69 | 69 | 69 | 67 | 67 |
| Years | 34 | 34 | 34 | 34 | 34 | 34 |
| Coefficients | 469 | 469 | 469 | 469 | 2 | 732 |
| Exporter × importer FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Exporter × year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Importer × year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Cross-border × year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Standard error clustering | Exp× Imp | Exp, Imp | Exp× year, Imp× year | Exp, Imp, year | Exp× Imp | Exp× Imp |

**Notes**: The table presents PPML regression results using the ETWFE estimator (equation (3)), for which the cohort-time-specific treatment effects were aggregated using equation (7) to obtain an aggregate treatment effect estimate. The dependent variable is exports which vary over the exporter-importer-year dimension. The 'Baseline' sample contains 69 countries, accounting for 98% of world exports. Standard errors in parentheses are clustered by country pair in columns 'Baseline', (4), and (5), exporter and importer in column (1), exporter-year and importer-year in column (2), and exporter, importer, and year in column (3). Columns (4) and (5) report RTA effects for agreements for which the number of provisions is below the median ($DEPTH_{ij,t} < P50$) or above the median ($DEPTH_{ij,t} >= P50$), respectively. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

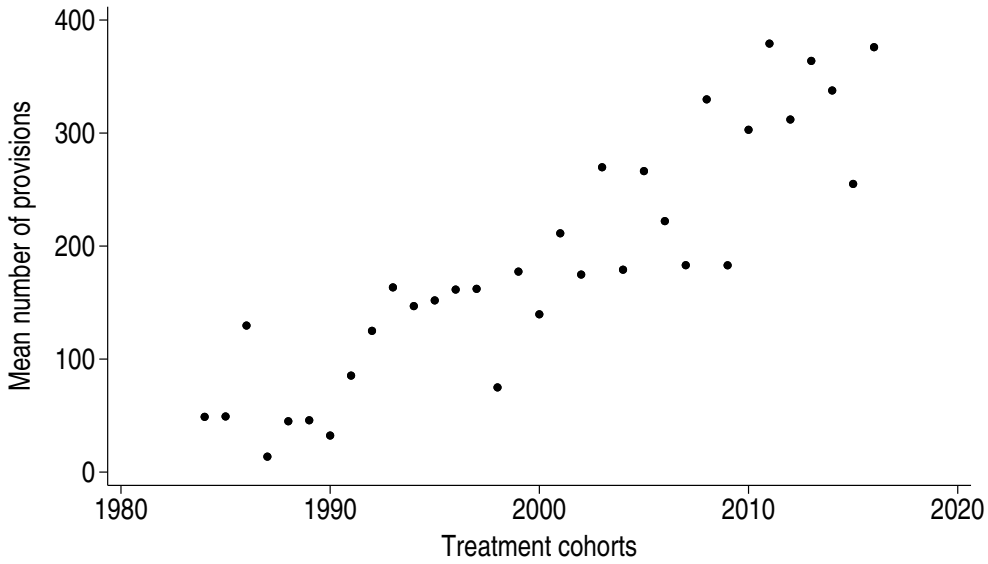
agreements include more provisions that are designed to shape international trade among their member countries. This is confirmed in Figure 8, where we plot the number of provisions in the agreements in our sample by cohort.[40] The general finding from the related literature is that 'deeper' agreements, i.e., those with more provisions, would lead to more trade among RTA members (Osnago et al. (2019) and Larch and Yotov (2023)).

Against this backdrop, the motivation for our next experiment is threefold. First, we want to check whether we can confirm that deeper agreements lead to more trade with the proposed ETWFE estimator. Second, we want to test whether our main result that the effects of RTAs are larger with the ETWFE estimator is confirmed for deep RTAs. Third, we want to demonstrate how the methods can be applied to study the impact of alternative agreement variables. Lastly, we want to explore whether and how we can reconcile our finding of falling RTA effects for more recent cohorts with the fact that more recent RTAs are deeper and, therefore, we would expect their impact to actually be

[40]Data on RTAs depth and number of provisions come from the World Bank's Database on the Content of Regional Trade Agreements (DCRTA) (Hofmann et al. (2019) and Mattoo et al. (2020)).

stronger rather then weaker. To keep the analysis simple, we split the agreements in our sample into two groups depending on whether the number of provisions that they include is above or below the mean for the sample.

Figure 8: Mean number of provisions by cohort




**Notes**: This figure shows the mean number of provisions per agreement by cohort, where we use the maximum in case the number of provisions changed over time for a given agreement. Data on the number of provisions comes from the World Bank's Database on the Content of Regional Trade Agreements (DCRTA) (Hofmann et al. (2019) and Mattoo et al. (2020)).

The results from this experiment appear in Table 10. The estimates in column (4) are obtained with the TWFE estimator, and they confirm that deep RTAs lead to greater trade liberalization. The estimates in column (5) are obtained with the ETWFE estimator and, based on those estimates, we conclude that the impact of deep RTAs is indeed stronger. In addition, comparing the results from columns (4) and (5) confirms our main findings that the ETWFE estimator delivers larger RTA estimates, both for the deep and more shallow agreements in our sample.

The overall conclusion that we draw from the battery of robustness experiments that we performed and described in this section is that they demonstrate that our main result that the ETWFE RTA effects are significantly larger than previously thought based on

TWFE estimates is confirmed and reinforced.

# 6 Conclusion

Motivated by the widely heterogeneous (across group and across time periods) policy estimates in the trade gravity literature and the concerns from recent econometric papers that TWFE estimates in staggered DiD designs, where the intervention occurs in multiple units in different time periods, may be biased, we nest an extended TWFE estimator *à la* Wooldridge (2023) within the empirical structural gravity model. To test the implications of the resulting methods, we estimate the impact of one of the most widely studied trade policies – regional trade agreements (RTAs).

The main results from our analysis are that the ETWFE estimator delivers RTA estimates that are significantly larger and longer lasting than corresponding estimates that are based on the current TWFE methods from the gravity literature. On a related note, the larger ETWFE estimates of the effects of RTAs are also more plausible from a policy perspective. Computing the implicit weights attached by the (OLS) TWFE to individual cohort-year effects following de Chaisemartin and D'Haultfœuille (2020) suggests that the TWFE is biased downward relative to the target parameter of the ETWFE estimate by putting larger weights on short-run effects and on those of late cohorts, which are both associated with smaller treatment effects. A series of sensitivity experiments confirm the robustness of our main findings, in particular, to the heterogeneity of the ETWFE estimator and potential IPPs, as well as to various specifications from the DiD and gravity literatures.

Based on our analysis of the impact of RTAs, we expect that the ETWFE estimator (and, in the future, related heterogeneity-robust staggered DiD methods generalized to the non-linear setting) may have significant implications for the estimates of other policies that have been studied with the gravity model of trade, e.g., economic sanctions, membership in GATT and WTO, currency unions, etc., which all share the common feature that the

treatment or intervention occurs in multiple units and in different time periods and are likely characterized by a large degree of treatment effect heterogeneity. Moreover, we expect that the new heterogeneity-robust staggered DiD methods may have significant implications for many of the policy estimates in gravity models of migration, FDI, and technology and innovation flows. Beyond the binary treatment considered in this paper, studying treatment effect heterogeneity for continuous treatments such as tariffs in the non-linear gravity settings is a promising area of future research.[41]

[41] See, for example, de Chaisemartin et al. (2022) for the linear case.

# References


**Agnosteva, Delina E., James E. Anderson, and Yoto V. Yotov.** 2019. "International Trade Costs: Assaying Regional Frictions." *European Economic Review* 112 32–50.

**Anderson, James E..** 2011. "The Gravity Model." *Annual Review of Economics* 3 133–160.

**Anderson, James E., Mario Larch, and Yoto V. Yotov.** 2018. "GEPPML: General equilibrium analysis with PPML." *The World Economy* 41 (10): 2750–2782.

**Anderson, James E., Mario Larch, and Yoto V. Yotov.** 2019. "Trade and investment in the global economy: A multi-country dynamic analysis." *European Economic Review* 120 103311.

**Anderson, James E., and Eric van Wincoop.** 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review* 93 (1): 170–192.

**Anderson, James E., and Yoto V. Yotov.** 2016. "Terms of Trade and Global Efficiency Effects of Free Trade Agreements, 1990-2002." *Journal of International Economics* 99 (C): 279–298.

**Athey, Susan, and Guido W. Imbens.** 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74 (2): 431–497, http://www.jstor.org/stable/3598807.

**Athey, Susan, and Guido W. Imbens.** 2022. "Design-based analysis in Difference-In-Differences settings with staggered adoption." *Journal of Econometrics* 226 (1): 62–79. https://doi.org/10.1016/j.jeconom.2020.10.012, Annals Issue in Honor of Gary Chamberlain.

**Baier, Scott L., and Jeffrey H. Bergstrand.** 2007. "Do Free Trade Agreements Actually Increase Members' International Trade?" *Journal of International Economics* 71 (1): 72–95.

**Baier, Scott L., and Jeffrey H. Bergstrand.** 2021. "NSF-Kellogg Institute Data Base on Economic Integration Agreements." *Kellogg Institute* https://kellogg.nd.edu/nsf-kellogg-institute-data-base-economic-integration-agreements.

**Baier, Scott L., Jeffrey H. Bergstrand, and Matthew W. Clance.** 2018. "Heterogeneous effects of economic integration agreements." *Journal of Development Economics* 135 587–608. https://doi.org/10.1016/j.jdeveco.2018.08.014.

**Baier, Scott L., Yoto V. Yotov, and Thomas Zylkin.** 2019. "On the Widely Differing Effects of Free Trade Agreements: Lessons from Twenty Years of Trade Integration." *Journal of International Economics* 116: 206–226.

**Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang.** 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144 (2): 370–395. https://doi.org/10.1016/j.jfineco.2022.01.004.

**Baldwin, Richard E., and Daria Taglioni.** 2006. "Gravity for Dummies and Dummies for Gravity Equations." Working Paper 12516, National Bureau of Economic Research.

**Bergstrand, Jeffrey H., Mario Larch, and Yoto V. Yotov.** 2015. "Economic Integration Agreements, Border Effects, and Distance Elasticities in the Gravity Equation." *European Economic Review* 78: 307–327.

**Borusyak, Kirill, and Xavier Jaravel.** 2017. "Revisiting Event Study Design." *Harvard University Working Paper*.

**Borusyak, Kirill, Xavier Jaravel, and Jann Spiess.** 2023. "Revisiting Event Study Designs: Robust and Efficient Estimation." Papers 2108.12419, arXiv.org, https://ideas.repec.org/p/arx/papers/2108.12419.html.

**Callaway, Brantly, and Pedro H.C. Sant'Anna.** 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001, Themed Issue: Treatment Effect 1.

**Carrere, Celine, Monika Mrazova, and J Peter Neary.** 2020. "Gravity Without Apology: the Science of Elasticities, Distance and Trade." *The Economic Journal* 130 (628): 880–910.

**Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer.** 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *The Quarterly Journal of Economics* 134 (3): 1405–1454. 10.1093/qje/qjz014.

**de Chaisemartin, Clément, and Xavier D'Haultfoeuille.** 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." Working Paper 29873, National Bureau of Economic Research. 10.3386/w29873.

**de Chaisemartin, Clément, Xavier D'Haultfoeuille, Félix Pasquier, and Gonzalo Vazquez-Bare.** 2022. "Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period." Papers 2201.06898, arXiv.org, https://ideas.repec.org/p/arx/papers/2201.06898.html.

**de Chaisemartin, Clément, and Xavier D'Haultfœuille.** 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–2996. 10.1257/aer.20181169.

**Cheng, I-Hui, and Howard J. Wall.** 2005. "Controlling for Heterogeneity in Gravity Models of Trade and Integration." *Federal Reserve Bank of St. Louis Review* 87 (1): 49–63.

**Ciani, Emanuele, and Paul Fisher.** 2019. "Dif-in-Dif Estimators of Multiplicative Treatment Effects." *Journal of Econometric Methods* 8 (1): 20160011. doi:10.1515/jem-2016-0011.

**Dai, Mian, Yoto V. Yotov, and Thomas Zylkin.** 2014. "On the Trade-Diversion Effects of Free Trade Agreements." *Economics Letters* 122 (2): 321–325.

**Deshpande, Manasi, and Yue Li.** 2019. "Who Is Screened Out? Application Costs and the Targeting of Disability Programs." *American Economic Journal: Economic Policy* 11 (4): 213–248. 10.1257/pol.20180076.

**de Chaisemartin, Clément, and Xavier D'Haultfœuille.** 2022. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey." *The Econometrics Journal*. 10.1093/ectj/utac017.

**Dhaene, Geert, and Koen Jochmans.** 2015. "Split-panel Jackknife Estimation of Fixed-effect Models." *The Review of Economic Studies* 82 (3): 991–1030. 10.1093/restud/rdv007.

**Eaton, Jonathan, and Samuel Kortum.** 2002. "Technology, Geography and Trade." *Econometrica* 70 (5): 1741–1779.

**Egger, Peter H, Marko Koethenbuerger, and Gabriel Loumeau.** 2021. "Local border reforms and economic activity." *Journal of Economic Geography* 22 (1): 81–102. 10.1093/jeg/lbab030.

**Egger, Peter H., Mario Larch, and Yoto V. Yotov.** 2022. "Gravity Estimations with Interval Data: Revisiting the Impact of Free Trade Agreements." *Economica* 89 (353): 44–61.

**Egger, Peter H., and Sergey Nigai.** 2016. "World-Trade Growth Accounting." *CESifo Working Paper No. 5831*.

**Egger, Peter H., and Filip Tarlea.** 2015. "Multi-way clustering estimation of standard errors in gravity models." *Economics Letters* 134 (C): 144–147.

**Egger, Peter, and Mario Larch.** 2008. "Interdependent Preferential Trade Agreement Memberships: An Empirical Analysis." *Journal of International Economics* 76 (2): 384–399.

**Felbermayr, Gabriel, Aleksandra Kirilakha, Constantinos Syropoulos, Erdal Yalcin, and Yoto V. Yotov.** 2020. "The global sanctions data base." *European Economic Review* 129 103561.

**Fernández-Val, Iván, and Martin Weidner.** 2016. "Individual and time effects in nonlinear panel models with large N, T." *Journal of Econometrics* 192 (1): 291–312. https://doi.org/10.1016/j.jeconom.2015.12.014.

**Gardner, John.** 2022. "Two-Stage Differences in Differences." *Working Paper*.

**Gurevich, Tamara, and Peter Herman.** 2018. "The Dynamic Gravity Dataset: 1948-2016." USITC Working Paper 2018-02-A.

**Heckman, James J., Hidehiko Ichimura, and Petra E. Todd.** 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64 (4): 605–654, http://www.jstor.org/stable/2971733.

**Hofmann, Claudia, Alberto Osnago, and Michele Ruta.** 2019. "The Content of Preferential Trade Agreements." *World Trade Review* 18 (3): 365–398.

**Hufbauer, Gary C., and Barbara Oegg.** 2003. "The Impact of Economic Sanctions on US Trade: Andrew Rose's Gravity Model." *Peterson Institute for International Economics*.

**Hummels, D.** 2001. "Toward a Geography of Trade Costs." unpublished manuscript, available for download at http:// www.krannert.purdue.edu/faculty/hummelsd/research/toward/TGTC.pdf.

**Jochmans, Koen.** 2017. "Two-Way Models for Gravity." *The Review of Economics and Statistics* 99 (3): 478–485. 10.1162/REST_a_00620.

**Larch, Mario, José-Antonio Monteiro, Roberta Piermartini, and Yoto Yotov.** 2019a. "On the Effects of GATT/WTO Membership on Trade: They are Positive and Large After All." School of Economics Working Paper Series 2019-4, LeBow College of Business, Drexel University, https://ideas.repec.org/p/ris/drxlwp/2019_004.html.

**Larch, Mario, Joschka Wanner, Yoto V. Yotov, and Thomas Zylkin.** 2019b. "Currency Unions and Trade: A PPML Re-assessment with High-dimensional Fixed Effects." *Oxford Bulletin of Economics and Statistics* 81 (3): 487–510.

**Larch, Mario, and Yoto V. Yotov.** 2023. "Estimating the Effects of Trade Agreements: Lessons From 60 Years of Methods and Data." School of Economics Working Paper Series 2023-4, Drexel University.

**Mattoo, Aaditya, Nadia Rocha, and Michele Ruta.** eds. 2020. *Handbook of Deep Trade Agreements*. https://openknowledge.worldbank.org/handle/10986/34055, Washington, DC: World Bank.

**Moser, Christoph, and Andrew K. Rose.** 2012a. "Why Do Trade Negotiations Take So Long?" *Journal of Economic Integration* 27 (2): 280–290.

**Moser, Christoph, and Andrew K. Rose.** 2012b. "Why Do Trade Negotiations Take So Long?" *VoxEU, June 8, 2012,* available at https://cepr.org/voxeu/columns/why-do-trade-negotiations-take-so-long.

**Moser, Christoph, and Andrew K. Rose.** 2014. "Who Benefits From Regional Trade Agreements? The View From the Stock Market." *European Economic Review* 68 31–47. 10.1016/j.euroecorev.2014.01.012.

**Olivero, María Pía, and Yoto V. Yotov.** 2012. "Dynamic Gravity: Endogenous Country Size and Asset Accumulation." *Canadian Journal of Economics* 45 (1): 64–92.

**Osnago, Alberto, Nadia Rocha, and Michele Ruta.** 2019. "Deep trade agreements and vertical FDI: The devil is in the details." *Canadian Journal of Economics* 52 (4): 1558–1599.

**Persyn, Damiaan, and Wouter Torfs.** 2016. "A gravity equation for commuting with an application to estimating regional border effects in Belgium." *Journal of Economic Geography* 16 (1): 155–175.

**Pfaffermayr, Michael.** 2021. "Confidence intervals for the trade cost parameters of cross-section gravity models." *Economics Letters* 201 109787. https://doi.org/10.1016/j.econlet.2021.109787.

**Rios-Avila, Fernando.** 2022. "JWDID: Stata module to estimate Difference-in-Difference models using Mundlak approach." Statistical Software Components, Boston College Department of Economics, August.

**Rose, Andrew K.** 2000. "One money, one market: The effect of common currencies on trade." *Economic Policy* 15 (30): 7–45.

**Rose, Andrew K..** 2004. "Do We Really Know That the WTO Increases Trade?" *American Economic Review* 94 (1): 98–114.

**Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe.** 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics* 235 (2): 2218–2244. https://doi.org/10.1016/j.jeconom.2023.03.008.

**Santos Silva, João M.C., and Silvana Tenreyro.** 2006. "The Log of Gravity." *Review of Economics and Statistics* 88 (4): 641–658.

**Santos Silva, João M.C., and Silvana Tenreyro.** 2021. "The Log of Gravity At 15." School of Economics Discussion Papers 0121, School of Economics, University of Surrey, `https://ideas.repec.org/p/sur/surrec/0121.html`.

**Schmidheiny, Kurt, and Sebastian Siegloch.** 2020. "On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization." ZEW Discussion Papers 20-017, ZEW - Leibniz Centre for European Economic Research, `https://ideas.repec.org/p/zbw/zewdip/20017.html`.

**Sun, Liyang, and Sarah Abraham.** 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006, Themed Issue: Treatment Effect 1.

**Teti, Feodora.** 2020. "30 Years of Trade Policy: Evidence from 5.7 Billion Tariffs." ifo Working Paper Series 334, ifo Institute - Leibniz Institute for Economic Research at the University of Munich, `https://ideas.repec.org/p/ces/ifowps/_334.html`.

**Tinbergen, Jan.** 1962. *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: The Twentieth Century Fund.

**Weidner, Martin, and Thomas Zylkin.** 2021. "Bias and consistency in three-way gravity models." *Journal of International Economics* 132 103513. https://doi.org/10.1016/j.jinteco.2021.103513.

**Wooldridge, Jeffrey M.** 2021. "Two-way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." *Available at SSRN 3906345*.

**Wooldridge, Jeffrey M.** 2023. "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data." *The Econometrics Journal*.

**Yotov, Yoto V.** 2022. "On the role of domestic trade flows for estimating the gravity model of trade." *Contemporary Economic Policy*, https://doi.org/10.1111/coep.12567.

**Yotov, Yoto V., Roberta Piermartini, José-Antonio Monteiro, and Mario Larch.** 2016. *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*. Geneva, Switzerland: United Nations and World Trade Organization, available for download at http://vi.unctad.org/tpa/index.html.