Supplementary Appendix

This appendix is not intended for publication. It accompanies "Staggered Difference-in-Differences in Gravity Settings: Revisiting the Effects of Trade Agreements" by Arne Nagengast and Yoto V. Yotov, and includes additional descriptive statistics (Subsection A.1), additional estimation results (Subsection A.2), and robustness checks (Subsection A.3).

A.1 Additional descriptive statistics

Table A1 reports the number of observations, country pairs, exporters, importers, and years for different groups in the baseline estimation sample from the ETWFE estimate in column (2) of Table 1. 'Cohort' refers to all post-treatment observations of country pairs with an RTA onset in a particular year. 'Treated' refers to all post-treatment observations of all cohorts. 'Not-yet treated' refers to all pre-treatment years of all cohorts. 'Never treated' refers to all observations of country pairs that did not sign an RTA agreement during the sample period.

Table A1: 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	10	27
1991 cohort	385	15	9	11	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	10	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

Table A2 reports the average of the variables Distance (in kilometers), Contiguity, Language, and Colony for different groups in the baseline estimation sample from the ETWFE estimate in column (2) of Table 1. '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.

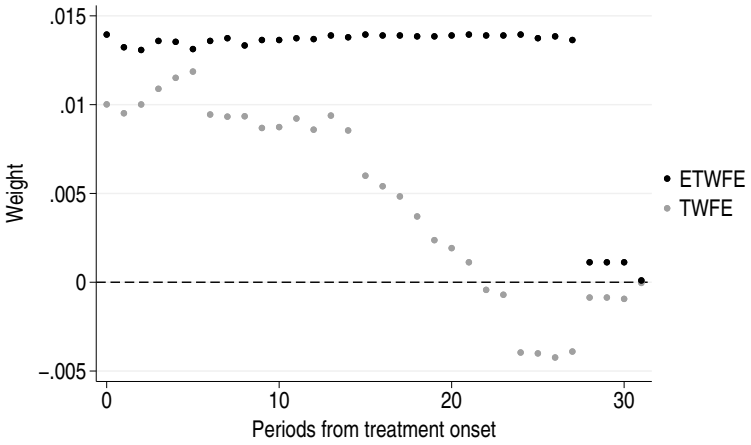
Table A2: Descriptive statistics: Summary statistics of covariates for different groups

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

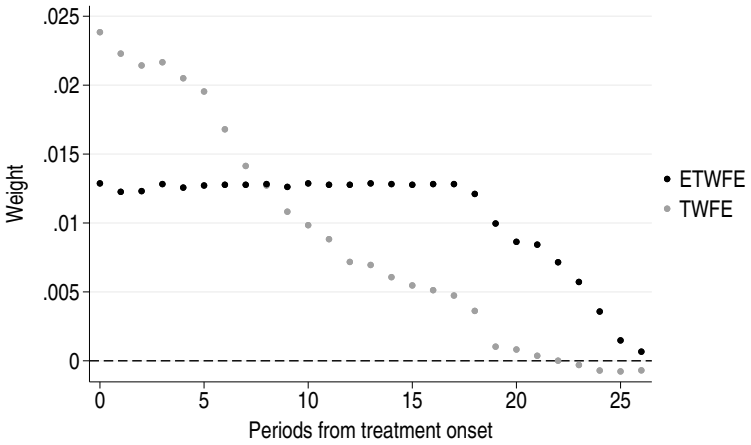
A.2 Additional results

Figure A1 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'). Panels (a)–(c) report weights aggregated by event time for the 1983-1989 cohort, the 1990-1999 cohort, and the 2000-2016 cohort, respectively. The figure shows that the cohort effect discussed in Section 4.4 is not mechanically driven by compositional differences in terms of event years since early (late) cohorts are more strongly underweighted (overweighted) in all years.

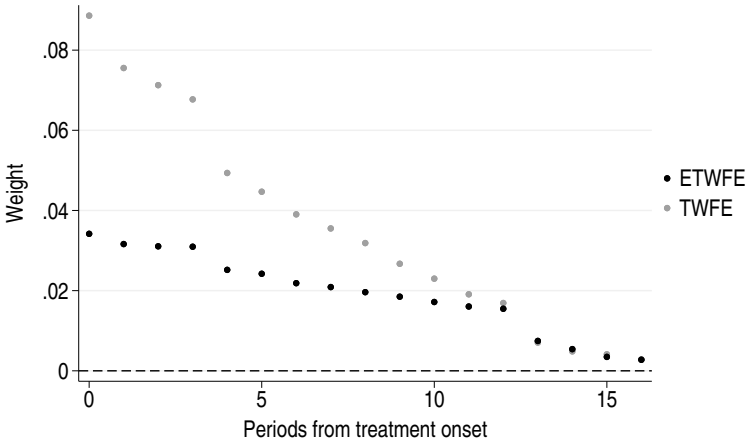
Figure A1: Weights of OLS ETWFE and TWFE estimator



(a) By event time (1983–1989 cohort)



(b) By event time (1990–1999 cohort)



(c) By event time (2000–2016 cohort)

A.3 Details on the robustness experiments

This section presents the results from a series of sensitivity experiments and robustness checks. To ease exposition and add some structure to the analysis, we group the experiments in four categories, which correspond to the four subsections of this section and cover (i) the degree of heterogeneity of the estimator, (ii) potential incidental parameter problems (IPPs), (iii) DiD-related experiments, and (iv) gravity-related experiments.

A.3.1 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 δ_{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 A3.

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 1), 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.

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 (Egger and Larch, 2008; Larch, 2021), while restricting the time heterogeneity to 5-year intervals for computational reasons.[A1] 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 obtains noisy (yet not consistent) estimates of individual treatment effects, which can be then be used to compute more aggregated average treatment effects (Borusyak et al., forthcoming). 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

[A1] 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.

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 by Wooldridge (2023).[A2]

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.[A3] 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.

[A2] While we do not report standard errors for the imputation estimator, they could likely be easily obtained using a bootstrap procedure.

[A3] 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. (forthcoming) 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.

Table A3: 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 A.3.2. '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.

A.3.2 Potential incidental parameter problems

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.[A4] The results are computed using 1,000 repetitions and displayed in Table A4.

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

[A4] For the simulation analysis, we assume the same data generating process as Weidner and Zylkin (2021), but we add cross-border $\times$ 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 β is set to 0.5, i.e., in the simulation analysis, we assume treatment effect homogeneity across cohorts and time.

biased.[A5]

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.

Table A4: 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 A.3.2 using 1,000 repetitions. Average bias refers to the mean of the difference between $\hat{\beta}$ and β. 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).

Jackknife bias correction. Motivated by the simulation results, we also apply the TWFE and ETWFE jackknife estimator in Table 1. In the baseline sample, 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),[A6] 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

[A5] For our baseline, we estimate 469 coefficients with 19,642 post-treatment observations in the treatment group (105,409 observations in total), i.e., around 42 (225) observations per coefficient.

[A6] 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).

standard error estimate is now larger than the 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.

For the medium and large sample, we obtain qualitatively similar results except that for the large sample the standard error of the ETWFE estimate remains slightly smaller than for the TWFE estimate. For the medium sample, a Z-test yields a p-value of 0.021, i.e., the null hypothesis of equality of coefficients is rejected at the 5% level. For the large sample, the corresponding p-value is 0.14, i.e., slightly above conventionally used levels of statistical significance.

As a result of these analyses, we recommend computing jackknife coefficient and standard error estimates for the ETWFE in the three-way gravity setting at least as a robustness check.

A.3.3 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 of our results 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.

Table A5: 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., forthcoming) 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.

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 A5. 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.

Alternative control groups. Next, we experiment with alternative control groups. Our findings are reported in Table A5. 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., forthcoming) 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 A5 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.

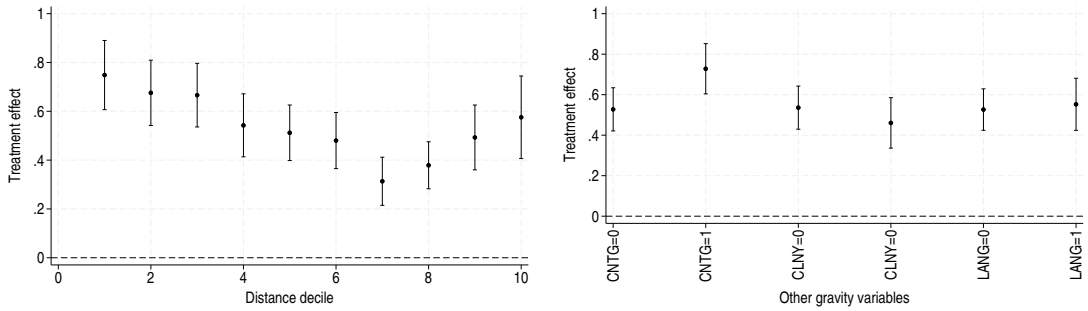
Time-constant covariates. Next, we consider an extension of the ETWFE estimation approach in which the cohort-time-specific treatment effects are allowed to vary by time-constant covariates (Wooldridge, 2023). As discussed in Wooldridge (2021), this is a parametric version of the regression adjustment 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 A2). 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 A2: Treatment effect by covariates



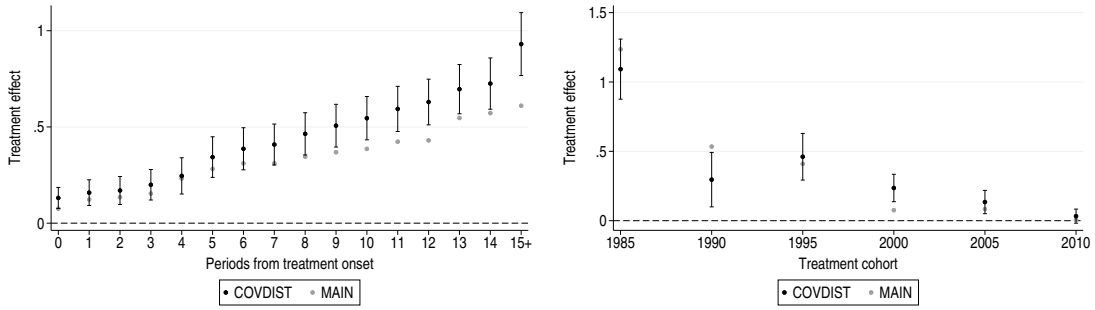
(a) By distance (b) By other gravity variables

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 A5. 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 A3). 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 A3: Event-time-specific and cohort-specific treatment effects



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

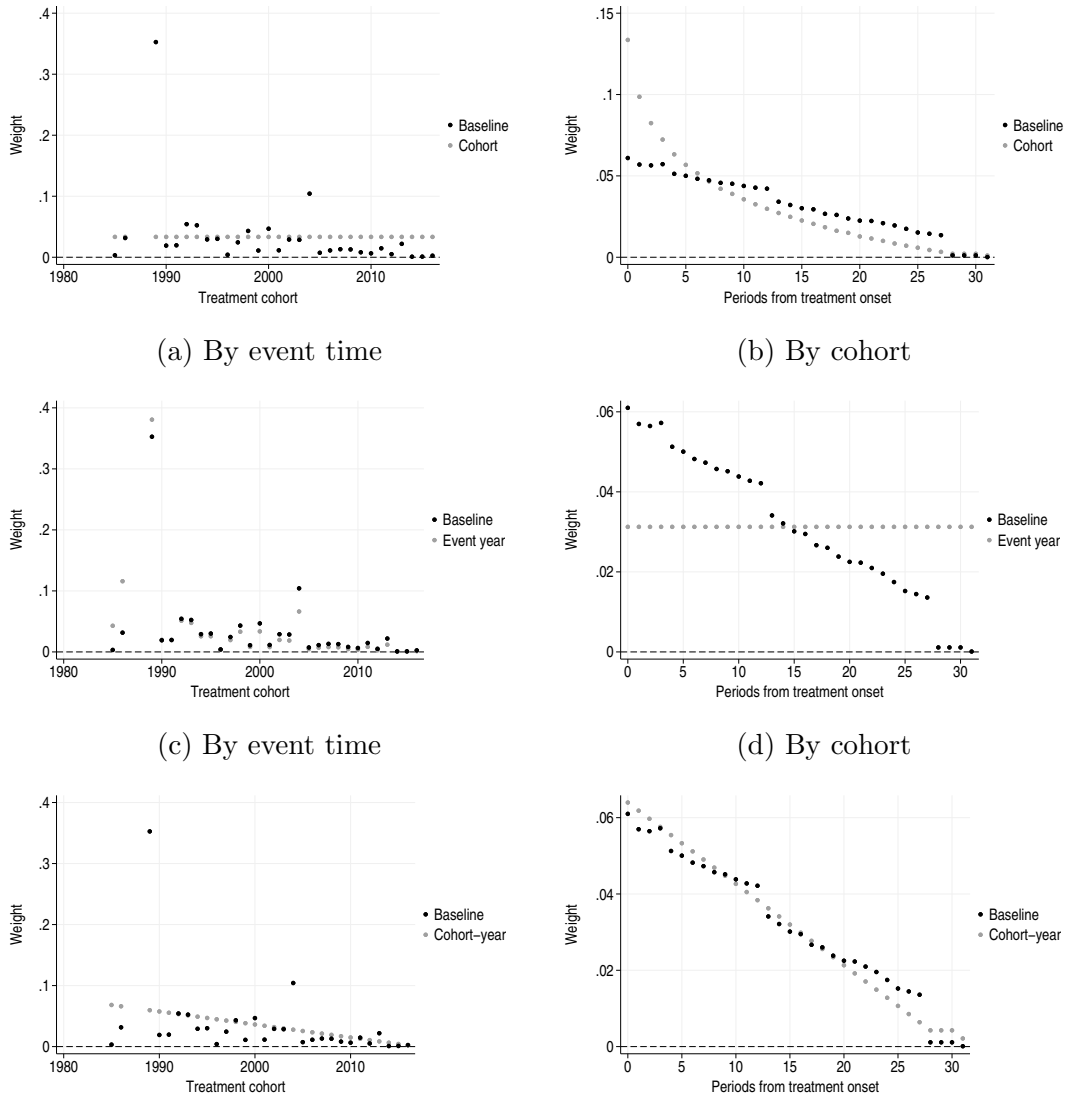
Notes: The figure reports different aggregations of the cohort-year-specific treatment effects from PPML estimation of equation (3). Panel (a) reports event-time-specific treatment effects from equation (3) aggregated using equation (9) from the covariate-augmented specification of column (8) in Table A5 in dark color ('COVDIST') along with the baseline specification of column (2) in Table 1 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 A5 in dark color ('COVDIST') along with the baseline specification of column (2) in Table 1 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 result-

ing weights by event time and by cohort are displayed in Figure A4 along with the weights of our baseline ETWFE estimate, and the corresponding aggregate treatment effects are reported in columns (10)-(12) of Table A5.

Figure A4: 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.

A.3.4 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 1. The main results from our gravity-related experiments are reported in Table A6.

OLS estimator. We start by reproducing our main results from columns (1) and (2) of Table 1 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., forthcoming; 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 A6: 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 A6. First, the TWFE estimate in column (1) of Table A6 is very close to the corresponding PPML result from column (1) of Table 1. Second, the ETWFE OLS estimate from column (2) of Table A6 is a bit smaller than the corresponding PPML result from Table 1, yet it is still more than twice as large as the TWFE OLS estimate from column (1) of Table A6. 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 A6. Even though the new TWFE estimate is a bit smaller than our main estimate from column (1) of Table 1, and the new ETWFE estimate is a bit larger than the corresponding estimate from column (1) of Table 1, 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.[A7] Therefore, in our next experiment, we only use data on international trade flows. The results are reported in columns (5) and (6) of Table A6. Two findings stand out. First, both the TWFE and the ETWFE estimates from Table A6 are significantly smaller than their counterparts from Table 1. 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 1, it is still significantly larger than the TWFE estimate from column (5) of Table A6, 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.[A8] 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 A6 and support our main conclusions. Specifically, we see

[A7] 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.

[A8] 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 A5. 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 1, 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 A6 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 1. 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 A6, 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 A7 (ETWFE) and Table A8 (TWFE), we reproduce our main results (i.e., of each of the three samples from Table 1) 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 A7: 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 A8: 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 A7 and A8 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.[A9]

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 A9. 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

[A9]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 A9: 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.

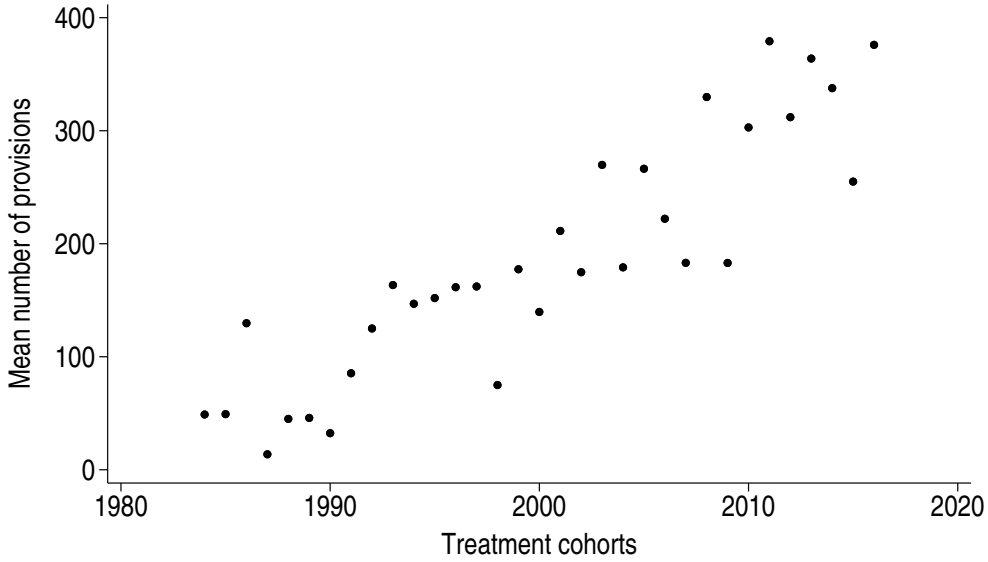
agreements include more provisions that are designed to shape international trade among their member countries. This is confirmed in Figure A5, where we plot the number of provisions in the agreements in our sample by cohort.[A10] The general finding from the related literature is that 'deeper' agreements, 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

[A10] 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; Mattoo et al., 2020; World Bank, 2021).

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 A5: 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; Mattoo et al., 2020; World Bank, 2021).

The results from this experiment appear in Table A9. 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.

Subgroup results and counterfactual. In our last experiment, we study the effects of the determinants of the heterogeneity of the RTA effects across cohorts.[A11] To do so, we take the estimated treatment effects $\hat{\delta}_{i,j,s}$ from the ETWFE OLS model of column (2) of

[A11] We thank an anonymous referee for suggesting the analysis in this subsection.

Table A6, where s references time since treatment, i and j reference country pair (i, j), and the treatment effect is the same for all country pairs in a given cohort. Then, we regress $\hat{\delta}_{i,j,s}$ on a set of s fixed effects, a series of bilateral gravity variables, as well as country-specific characteristics. For time-varying variables, we use values for the years prior to treatment onset. To account for the fact that $\hat{\delta}_{i,j,s}$ is estimated, for inference, we rely on a bootstrap procedure sampling (i, j) pairs. From a methodological perspective, this analysis is in the spirit of de Chaisemartin and D'Haultfoeuille (2022) and Shahn (2023). From a policy perspective, it resembles Baier et al. (2019), who study the determinants of the effects of FTAs that are obtained from a TWFE estimation.

The estimation results are presented in Table A10, in which we consecutively introduce more explanatory variables at the cost of sample size due to missing values, while s fixed effects are always controlled for. The specification in column (1) considers a set of standard gravity variables. Consistent with Figure A2a and the previous literature (Baier et al., 2018, 2019), we find a negative impact of distance on the RTA effects on trade. Furthermore, we obtain a positive coefficient on language and a negative coefficient on previous colonial relationships, while the coefficient on contiguity is not statistically different from zero. Following Baier et al. (2019), in column (2), we also add the estimated pair fixed effect from the first stage (i.e., the specification in column (2) of Table A6) to control for the level of trade frictions between i and j before the signing of the corresponding RTA. However, this leaves the results almost unchanged, while the corresponding coefficient is negative as in Baier et al. (2019). In column (3), we add the log of the GDP of the exporter and importer as proxies for market size. In contrast to Baier et al. (2019), however, the resulting coefficients are not statistically significant. In column (4), we additionally add the log of applied tariffs, which is again not statistically significant. In column (5), we consider quintiles of distance and tariffs to account for potential non-linearities. In line with Figure A2a, we find evidence for a non-linear effect of distance on the effectiveness of RTAs. The tariff results suggests that higher tariffs (apart from the highest quintile) might be associated with higher trade effects in line with the intuition that higher pre-

treatment tariffs might leave more scope for the beneficial effects of trade liberalization (Baier et al., 2019), while the results do not reach statistical significance.

Finally, building on the subgroup-specific results, we try to answer the following counterfactual question: *"What would be the increase in bilateral trade if all country pairs not currently linked by an RTA were to sign one?"* From a policy perspective, this counterfactual is particularly important given the insignificant RTA effects for late-treated cohorts we found in Section 4.3. To answer this question, we rely on our estimates from columns (3) or (4) of Table A10 and values for the corresponding variables for all untreated country pairs, and we use them to predict the trade volume effects if the untreated countries were to sign RTAs with each other.[A12]

For the specification in column (3), we find an average coefficient of 0.076 after 5 years (increase of 7.9% in bilateral trade), 0.166 after 10 years (increase of 18.1% in bilateral trade), and 0.237 after 15 years (increase of 26.7% in bilateral trade) based on 2,185 country pairs for which the necessary covariate data is available. Similarly, for the specification in column (4), we find an average coefficient of -0.022 after 5 years (decrease of 2.2% in bilateral trade), 0.079 after 10 years (increase of 8.2% in bilateral trade), and 0.162 after 15 years (increase of 17.6% in bilateral trade) based on 643 country pairs. Therefore, the counterfactual results provide tentative evidence that the RTAs between remaining country pairs would significantly boost trade – in contrast to the results for late-treated cohorts – albeit to a lesser extent than the historical average in the baseline sample.

[A12]We are keenly aware that this is simply a partial equilibrium counterfactual analysis and, to obtain the full RTA effects, one would need to take into account possible general equilibrium trade diversion effects, which will mitigate our partial equilibrium predictions.

Table A10: Subgroup-specific results

	(1)	(2)	(3)	(4)	(5)
$\ln DIST_{ij}$	-0.563***	-0.662***	-0.748***	-1.183***	
	(0.086)	(0.086)	(0.104)	(0.328)	
$CNTG_{ij}$	0.030	0.054	0.027	0.114	0.218
	(0.052)	(0.053)	(0.057)	(0.250)	(0.249)
$CLNY_{ij}$	-0.225***	-0.201***	-0.206***	-0.169	-0.379
	(0.079)	(0.075)	(0.072)	(0.188)	(0.264)
$LANG_{ij}$	0.082***	0.092***	0.136***	0.124	0.119
	(0.020)	(0.021)	(0.023)	(0.077)	(0.073)
1st-stage pair FE		-0.016***	-0.026***	-0.018	-0.016
		(0.004)	(0.007)	(0.022)	(0.021)
$\ln GDP_i$			0.008	-0.018	-0.023
			(0.011)	(0.038)	(0.037)
$\ln GDP_j$			0.004	-0.024	-0.038
			(0.010)	(0.035)	(0.035)
$\ln TARIFF_{ij}$				-0.148	
				(0.357)	
$DIST_{ij} = Q2$					-0.341**
					(0.144)
$DIST_{ij} = Q3$					-0.435***
					(0.148)
$DIST_{ij} = Q4$					-0.429***
					(0.140)
$DIST_{ij} = Q5$					-0.300**
					(0.126)
$TARIFF_{ij} = Q2$					0.025
					(0.091)
$TARIFF_{ij} = Q3$					0.148
					(0.137)
$TARIFF_{ij} = Q4$					0.095
					(0.106)
$TARIFF_{ij} = Q5$					-0.059
					(0.111)
Observations	19,574	19,574	18,294	7,465	7,465
Exporters	66	66	66	40	40
Importers	66	66	66	40	40
Years	32	32	32	28	28
Event year FE	Yes	Yes	Yes	Yes	Yes

Notes: The table presents OLS regression results of the estimated treatment effects $\hat{\delta}_{i,j,s}$ from the ETWFE OLS model of column (2) of Table A6 on a set of s fixed effects to control for dynamics and country pair as well as country characteristics. For time-varying variables, we use values for the years prior to treatment onset. The 1st-stage pair FE are from the regression in column (2) of Table A6. $\ln TARIFF_{ij}$ refers to the natural log of 1 + the ad valorem applied tariff between i and j from Baier et al. (2019). $DIST_{ij} = Q2$ ($TARIFF_{ij} = Q2$) refers to the second quintile of distance (tariffs) in the estimation sample etc. Standard errors in parentheses are obtained using a bootstrap procedure sampling (i, j) pairs. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.