A match made in Maastricht: Estimating the treatment effect of the euro on trade

Joseph Kopecky*

31st October 2022

Abstract

Why do estimates of the European Monetary Union (EMU) effect on trade vary so greatly? Rose (2017) shows that the largest factor determining the size of EMU trade estimates is the choice of sample, with studies using only European or rich countries finding smaller impacts than those using more complete trade datasets. I push this question one step further, asking instead: what is the appropriate comparison group with which to study the euro's trade impact? Using a first stage estimation of selection into the EMU and a robust propensity score weighting estimator, I show that gravity estimates of the euro effect on trade are smaller when sample restriction and weighting brings the differences in observable characteristics between EMU and non-EMU pairs close to zero. Utilizing a Poisson pseudo-maximum likelihood approach, I find that estimates using this more robust estimator reflect the same pattern, but with significantly less initial upward bias. My work suggests that policy analysis in trade should be more careful to consider the comparability of "treated" and "control" observations, and more readily utilize propensity score methods as a data driven approach to re-balancing samples when differences across these groups are large.

*Trinity College, Dublin (jkopecky@tcd.ie).

1. Introduction

The euro represents an unprecedented policy experiment. There is little historical precedent for such a large number of wealthy countries to multilaterally surrender control of their currencies. In their assessment of the challenges facing the eurozone in the depths of the European sovereign debt crisis, O'Rourke and Taylor (2013) show that most prior currency unions make for a poor comparison to the European Monetary Union (EMU). It should not then be surprising that estimates of the impact of the EMU on trade differ so dramatically from those studying earlier currency unions. Assignment into the EMU, to borrow language from experimental studies: the policy *treatment*, is not remotely random. An important consideration is how accurately the *control* group (non-EMU country-pairs) compares on observable characteristics to those that receive policy treatment. In recent work surveying the literature on EMU estimates on trade, Rose (2017) shows that most of the small estimates of the EMU effect on trade come from sample reductions, advocating that more data should be better. I show that sample selection is crucial to determining the size of EMU coefficients, but that studies which reduce the sample[1] may at times offer an improvement in the comparability of these groups. I propose a data driven, propensity score, method of rebalancing the EMU and non-EMU samples to better match on observed characteristics. This not only provides an empirically driven way to select the sample, but also utilizes a robust weighting estimator that leverages gains from matching estimators, while still using well studied, and theoretically grounded, gravity equations of trade to estimate treatment effects.

The effect of currency unions on trade has been hotly debated. Empirical work stems from Rose (2000), whose gravity equation estimates suggest that a common currency more than doubles bilateral trade between countries. This was, by the author's own admission, unreasonably large but surprisingly robust (Rose, 2002). A cottage industry sprung up to minimize the currency union effect with Nitsch (2002) suggesting a weaker effect, and Glick and Rose (2002) showing that a smaller, though still extremely large estimate survives limiting estimation to the time-series *within-pair* variation. More recently attention has focused on estimating these effects for the euro-area, with Glick and Rose (2016) providing an update on their earlier currency union estimates to include the reasonably large time-series of euro data. They find an estimate of a roughly 54% increase in trade from EMU membership. This result is a much larger trade effect than much of the literature. A meta analysis carried out in Polák (2019) finds estimates of an EMU trade effect in the 2%-6% range, with more recent evidence pointing to little, if any effect at all. Rose (2017) suggests

[1] For example, by including only high income countries

that much of the difference among EMU trade estimates comes from sample selection, with effects that are increasing as new years are added to the sample, and smaller when limiting the sample to subsets of rich countries (roughly 12% gains) rather than the full bilateral export data as in Glick and Rose (2016).

Larch, Wanner, Yotov, and Zylkin (2019) reproduce the results of Glick and Rose (2016) using the PPML estimators suggested in Silva and Tenreyro (2006) and Silva and Tenreyro (2011), who show that log specifications of the gravity produce biased estimates in the presence of heteroscedasticity. Larch et al. (2019) show that this heteroscedasticity is particularly important in the context of EMU trade estimates. Notable for my work, they find that inclusion of small countries in the dataset produces sizeable impacts on estimates of the EMU effect, amplifying the difference between PPML and OLS results. They show, similar to Rose (2017), that estimates of the currency union and EMU effects vary substantially when altering the sample (OECD, Upper income, etc). My results will echo this conclusion, while contributing to their findings by providing a data driven method of sample selection, chosen to reduce the differences in observable characteristics between treated and control sub-populations. In addition to this, I find that the PPML estimator provides much more stable estimates when the underlying sample changes, further motivating its use relative to log-gravity estimators when problems of sample selection are large.

I am not the first to apply propensity score methods to the currency union effect on trade, nor even to the case of the EMU. Persson (2001) shows that using matching estimators to estimate the average treatment effect of currency unions significantly reduces their estimated impact. Similar matching estimators are used by Chintrakarn (2008) in the case of the euro area, again reducing estimators. Their work relies on estimation of selection into *treatment* groups, using logit and probit models of probability of being in a currency union (or the EMU) and then comparing the conditional mean of treated observations with that of control groups with similar likelihood of being treated. While these estimates work to solve the selection problem that is endemic in macro policy estimates of trade, they also fail to leverage the usefulness of the well studied gravity equations used in work such as Glick and Rose (2016). I instead use a *doubly robust* estimator, combining the propensity score weighting used in work such as Persson (2001) and Chintrakarn (2008) with calculations of the conditional mean using a gravity equation approaches that are standard to the trade literature. This and other doubly robust estimators, which model both selection into treatment as well as the policy effect on outcomes are described in detail in: Imbens (2004), Lunceford and Davidian (2004), and Wooldridge (2007); with applications in the context of macroeconomic policy (fiscal shocks) in Jordà and Taylor (2016).

To my knowledge the only other study that applies my preferred estimator, inverse propensity score weighting with regression adjustment (IPWRA), to the EMU is Millimet and Tchernis (2009). Their work is primarily interested in providing guidance on specification of the first-stage model, showing that overfitting such models can be beneficial. Their EMU application is applied to the period of 1999-2002 and focuses only on 22 developed countries. Their results find an impact in line with similar panel OLS estimates such as Micco, Stein, and Ordoñez (2003), finding a roughly 12% impact of EMU membership on log trade. My estimation updates their results to a much larger trade dataset, while contextualizing the use of the IPWRA as a method of improving on the problem of selection into EMU membership. In particular, this propensity score model provides an explanation for the observations made in Rose (2017), suggesting that the differences due to sample size may favor the smaller estimates on truncated trade data. Work such as Micco et al. (2003), and many others, restrict their estimation sample to developed countries. Generally these estimates rely on an *ad hoc* selection of sample. I show that doing so appears to improve the comparability of treatment and control observations (ie reduces the differences in means of observable characteristics), but by less than my preferred method, which provides a data driven approach to choosing an appropriate comparison sample.

Rose (2017) makes the case that more data should be better, suggesting that the omission of poor and small countries from the sample biases downward the EMU effect. However, problems of selection bias in estimates of currency unions are well known. It was precisely this issue that motivated Persson (2001) to use matching estimators. While the Rose (2001) makes compelling arguments for why the pure matching approach has flaws, the estimator I suggest utilizes the strengths of both these propensity scores and traditional gravity approaches. Knowing the potential issues of selection, increasing the sample size to include observations that have little in common with EMU members appears to bias estimates upwards. This works in the same way that including large pools of healthy, low risk, individuals as a comparison group may bias observational estimates on the efficacy of a medical treatment downward. Their better health makes them less likely to receive the treatment, but lack of treatment has no causal bearing on their health outcomes. The experimental ideal would compare those receiving the medical treatment to individuals who need such treatment, but do not receive it. Of course this is not possible with observational data. It is precisely such contexts for which the IPWRA style estimators (and earlier matching estimators) were developed.[2] While I do not claim that this estimator removes all potential issues of selection in the context of the EMU, improving the comparability of

[2]See for an example using IPWRA to study the effectiveness of right heart catheterization, Hirano and Imbens (2001)

treatment and control observations should only work to reduce such problems as much as possible given the observable characteristics available in standard bilateral trade data.

2. Data and methodology

I use the CEPII gravity dataset from Conte, Cotterlaz, and Mayer (2021), constructing a measure of EMU currency union membership consistent with existing measures from Glick and Rose (2016), but excluding some small French territories that are included in their analysis. In all results below, I drop non-EMU currency union pairs to ensure that my baseline comparison group is not polluted with countries currently in a different currency union pair that may similarly affect their trade. Comparing to Rose (2017), my full dataset includes more years, but fewer country-pairs, when restricting to log-exports (29,394 instead of 34,104), this is in part due to dropping non-EMU unions, and in part because of differences in data as Glick and Rose (2016) use IMF DOTS rather than Conte et al. (2021). As a result, my full sample estimates are close-to, but slightly different from the Glick and Rose (2016) result.

My starting point for estimation is the gravity specification suggested in Head and Mayer (2014). This is the "theory-consistent" method of specifying this empirical relationship, accounting for the multilateral resistance terms that are important in structural gravity equations. To accomplish this I include a full set of exporter-year and importer-year dummy variables. In addition to these, I include a time-invariant set of country-pair fixed effects, which have been shown by many to be important in the context of currency unions, and combined make up the preferred specification in Glick and Rose (2016). This is given by:

$$ln(X_{ijt}) = \gamma EMU_{ijt} + \beta RTA_{ijt} + \lambda_{it} + \psi_{jt} + \phi_{ij} + \epsilon_{ijt} \tag{1}$$

where X_{ijt} are exports from country i to country j at time t, EMU_{ijt} is a dummy variable representing a EMU membership for a country-pair in year t, RTA_{ijt} is control for regional trade agreements, λ_{it} exporter-time fixed effects, ψ_{jt} importer-time fixed effects, and ϕ_{ij} time-invariant country-pair fixed effects. While I can include a large set of standard gravity controls, nearly all are inestimable while using exporter/importer-year and pair fixed effects, so I omit them from discussion here. Including the few with enough variation to remain in this specification have no bearing on estimates of $\hat{\gamma}$.

Silva and Tenreyro (2006) provide a now well known critique of Equation 1, showing that under heteroskedasticity, the log-linearized model of trade leads to biased estimates. They show in both Silva and Tenreyro (2006) and Silva and Tenreyro (2011) that estimations of Equation 1 on trade data suffer substantially from this bias, suggesting instead to

use a Poisson pseudo-maximum likelihood (PPML) estimator without taking logs. This methodology has the advantage of not only dealing with the bias introduced by the log-linear approximation, but also allows for inclusion of zero trade flows. While I will continue to use the log gravity specification to link my work to the results of Rose (2017). I also provide estimates of the equivalent PPML estimation, of:

$$X_{ijt} = e^{\gamma EMU_{ijt}+\beta RTA_{ijt}+\lambda_{it}+\psi_{jt}+\phi_{ij}} + \epsilon_{ijt} \tag{2}$$

where controls and fixed effects are defined identically to those above.

2.1. Inverse Propensity Score Weighting: A Doubly Robust Estimator

Propensity scores feature heavily in the rigorous debate around the size of the currency union effect on trade. In his original rebuttal of the eye-popping estimates of Rose (2000), Persson (2001) showed that matching estimators substantially reduce the impact of the currency union effect on trade. His methodology uses two estimators: *nearest-neighbor* matching, and *stratification*. Both are two-step estimators that rely on a first stage estimate of the probability of selection into a currency union, and then generate estimates for the average treatment effect using a difference in means among these groups. The nearest neighbor method pairs each treated observation with its most comparable control group, while stratification bins treated and control observations according to their probability of treatment and takes a weighted difference in means within each bin. The goal of these methods is to deal with concerns of selection, well understood in the context of currency unions and the euro area.[3] Similar methods are used in a more recent estimate of the euro area in Chintrakarn (2008), who again suggests a downward revision of eurozone estimates after applying matching estimators.

In an initial response to Persson (2001), Rose and Honohan (2001) correctly critiques the probit model used for selection into currency unions as a poor fit for the data. Histograms of treated (currency union) probabilities reveal that most of the weight of predicted liklihood of being in a currency union falls near bottom of the distribution. Matching estimators such as those used in Persson (2001) rely on the identification of the model of selection into treatment for the second stage difference-in-means estimators. Identification hinges on converting observational data into something closer to a randomized control trial under the assumption that the rebalancing of treatment and controls will ameliorate selection issues. However, the lack of fit and out-of-sample predictions of currency unions should give

[3]See for example, Baldwin and Taglioni (2007) who discuss this issue in the context of estimating the euro effect on trade and Baldwin (2006) who compares this problem to an attempt to study the impacts of dieting on weight gain without understanding the underlying model whereby individuals decide to go on a diet.

caution to those relying on the soundness of this identification, especially when comparing their results to the well studied, and theoretically motivated, gravity equation. The case that Rose and Honohan (2001) makes against their use is rather convincing, particularly when comparing to the modern specifications of gravity that control for time-varying country-specific multilateral resistance terms, as Glick and Rose (2016) does.

However, such critiques of the chosen matching estimators do not address the concerns of selection that motivated the work of Persson (2001) and Kenen (2002). Perhaps it is possible to have our cake and eat it too? A different class of estimators use propensity score matching to estimate *doubly robust* estimators, that replace the difference-in-means second stage with other second stage estimates of conditional means. I make use of inverse propensity score weighting with regression adjustment (IPWRA). This process involves specifying a first-stage treatment estimation of selection into treatment (the EMU), but instead of a difference in means approach, I then estimate a second stage model using regression specifications with propensity scores serving as weights. This allows for estimates to be rebalanced using first stage estimates similar to that of Persson (2001), while still using the well studied gravity equations to estimate conditional mean across treated and control sets. The *doubly-robust* nature of the IPWRA estimator, is shown and discussed in great detail in work such as Imbens (2004), Lunceford and Davidian (2004), and Wooldridge (2007), and suggests that the estimator will yield consistent estimates of the average treatment effect if *either* model is correctly specified. I follow similar notation to Jordà and Taylor (2016), who develop this estimator in a macroeconomic context.[4] The IPWRA estimator is given by:

$$\widehat{ATE}_{IPWRA} = \frac{1}{n_1^*}\sum\left[\frac{EMU_{ijt}\left(m_1\left(Z_{ijt},\hat{\beta}\right)\right)}{\hat{p}_{ijt}}\right] - \frac{1}{n_0^*}\sum\left[\frac{\left(1-EMU_{ijt}\right)\left(m_0\left(Z_{ijt},\hat{\beta}\right)\right)}{1-\hat{p}_{ijt}}\right] \quad (3)$$

where EMU_{ijt} is a dummy representing whether a country-pair is in the eurozone at time t, and $\hat{p}_{ijt}$ is the first-stage estimate of probability of treatment. A general term for any second stage estimate of conditional means for treated (1) and control (0) groups is given by: $m_{0/1}\left(Z_{ijt},\hat{\beta}\right)$.[5] I estimate this conditional mean using Equation 1 or Equation 2, where $\hat{\beta}$ now respresents the vector of all estimated regression coefficients. Estimates of $\hat{\beta}$ can in principle be estimated separately for treated and control populations, as described Słoczyński and Wooldridge (2018), however this is not possible while using the

[4]They use local projections to study fiscal austerity shocks rather than the static gravity equation I employ here.

[5]To condense notation I refer to the control set as simply Z_{ijt} with $\hat{\beta}$ referring to all estimated parameters, this includes the rich set of fixed effects in Equation 1.

fully specified theory consistent gravity equation described by Head and Mayer (2014), as specifying the conditional mean across treatment and controls requires a high dimension of fixed effects, larger than the number of EMU observations in the sample. As such, I am limited in this application to the assumption that coefficients for estimation of these conditional means are identical, which is of course also used in traditional gravity specifications. In keeping with suggestions of made in Hirano and Imbens (2001) and Imbens (2004) these weighted averages are normalized, with $n_1^* = \sum \frac{EMU}{\hat{p}}$ and $n_0^* = \sum \frac{1-EMU}{1-\hat{p}}$, to ensure that the probability weights are normalized to sum to one.

2.2. Modeling Selection Into Currency Unions

I estimate selection into currency unions using a logistic specification controlling for a wide range of country and pair specific controls. I include in these a number of standard gravity equation variables, as well as other characteristics that may be important for the formation of bilateral currency union agreements. My first stage specification is given in Equation 4.

$$\begin{aligned} EMU_{ijt} = \theta_0 + \theta_1 \ln Y_{it} \times Y_{jt} + \theta_2 \ln y_{it} \times y_{jt} + \theta_3 \ln Dist_{ijt} \\ + \theta_4 \ln |Y_{it} - Y_{jt}| + \theta_5 \ln |y_{it} - y_{jt}| + \theta_6 |g_{it} - g_{jt}| \end{aligned} \tag{4}$$

In Equation 4, in addition to the first three terms, which are standard gravity equation estimates of size of output ($Y_{it/jt}$), incomes ($y_{it/jt}$), and distances[6] ($Dist_{ijt}$) between the two countries, I also include the differences in output, per-capita GDP, and GDP growth rates ($g_{i/j}$). These are important because the standard gravity terms may do a poor job capturing the fact that many trading partnerships occur between relatively rich and poor (or large and small) economies, in ways that may be systematically different for the average eurozone economy. Millimet and Tchernis (2009) show via Monte Carlo simulation that there are potential benefits of *over-specifying* the propensity score estimator, so I include squared terms of each of these regressors along with the linear terms listed in Equation 4.

3. First Stage Estimation and Sample Selection

To generate average treatment effects of the eurozone given by the IPWRA estimator of Equation 3, I must first specify the probability of selection into the EMU given by a first stage estimation of Equation 4. Because the model of selection into the eurozone is not my outcome of interest, and the control set (including all linear and squared terms) is large,

[6] I use the population weighted distance between the two most populous cities in each country in thousands of miles.

I do not report the regression outcomes from these first stage specifications here. Their coefficients are sensible when considering euro pairs relative to non-euro country pairs. For example, the model generally finds them to have positive coefficients for log product of GDP and GDP per capita, suggesting they have larger output and are richer than the average pair in the sample, but also negative terms for their absolute differences, which suggests they are closer in relative economic size and well being than the average trading partners. I report various model fits and properties of the predicted probabilities from a number of first stage estimates in Table 1.

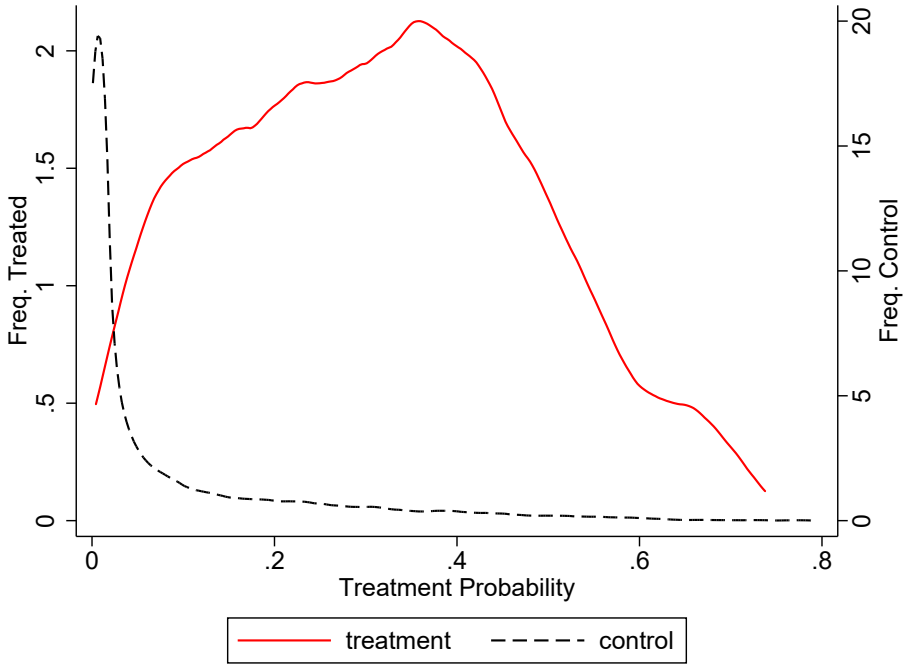
Table 1: *First Stage Summary Information*

	Probit	Logit	Probit	Logit
Controls (linear terms)	✓	✓	✓	✓
Controls (squared terms)			✓	✓
Observations	1,525,572	1,525,572	1,525,572	1,525,572
Pseudo-R^2	0.6366	0.6263	0.6793	0.6767
p-score range (EMU=1)	[0.00026, 0.9029]	[0.0008, 0.9272]	[0.0001, 0.7774]	[0.001, 0.811]
p-score mean (EMU=1)	0.28	0.28	0.31	0.32
p-score = 0 & Trade != 0	10,473	0	532,948	109,296
# non-zero p-score < min p-score (EMU=1)	1,400,706	1,421,544	217,422	612,177

All controls have p-values < 0.01*. Control set includes the log product of GDP and GDP per capita in origin and destination countries, log absolute differences in origin and destination GDP and GDP per capita, absolute difference in their output growth rates, population in both origin and destination, and the population weighted distance between largest cities.*

There is little difference in the model fit and mean treated probability in logit and probit specifications. Because fit is improved and Millimet and Tchernis (2009) suggests that IPWRA estimators may perform better with an *over-specified* model I will choose to use probabilities estimated using both the linear and squared terms of the variables in Equation 4. With such large data, and euro membership quite rare, the model struggles to fit this first stage particularly well, with the mean EMU member still having probability of joining at just 32%. This is a flaw of this approach, and why relying entirely on the strength of the matching estimator, as in work such as Persson (2001) and Chintrakarn (2008) has potential drawbacks. The strength of the IPWRA estimator is that I can attempt to rebalance treatment and control populations, but still use the gravity equation in Equation 1 or Equation 2 to calculate the conditional means. In all analysis below, I will use the logit estimator. This is because it has similar properties to the probit in terms of model fit, but with the fully saturated model many observations in the probit are assigned exactly zero and therefore their observations will be dropped in an estimation of Equation 3. These are

Figure 1: *Overlap: K-density plots of treated and control propensity scores (logistic model)*



Graph shows kernel density plots of treatment and control sub-populations with propensity scores >= the minimum for the treated group.

assigned extremely small, but positive, values by the logit model. Since I wish to compare my estimators to the log gravity estimates of Rose (2017), I keep baseline weighted estimate that is relatively close in sample size. As I will show below, the estimates on my sample with non-zero predicted propensity scores in the fully saturated logit will be similar to those in Rose (2017), making for a convenient baseline comparison for the rest of my results.

It is important that there is sufficient overlap in propensity score estimates, such that there are comparison observations across treatment and control groups. While the majority of weight of control population is near zero, I show in Figure 1 that there is overlap across the full distribution of treated observations. Although I use separate axes for readability, there are more non-EMU observations with a probability greater than the mean of 0.32 than EMU observations.

Truncating the sample to only include observations along the $[0.001, 0.811]$ distribution of the EMU observations results in a sample of $40,407$ (there are 4,690 pair-year observations between euro members). Such truncation is commonly used to limit the impacts of outliers. As can be seen in Equation 3, EMU observations that are predicted to be very likely to be non-EMU and non-EMU observations with high predicted probability of being in the EMU can receive high weights at the extreme tails of the probability distribution. Many estimates

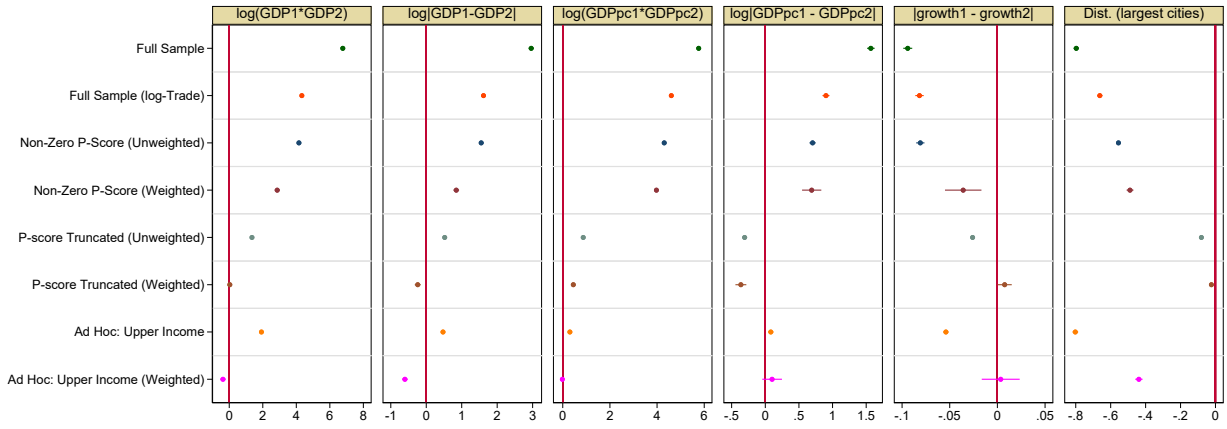
making use of IPWRA and other propensity score estimates consider truncating the sample to remove such outliers. This often done using an arbitrary rule-of-thumb method such as 1% or 5% cutoffs. Imbens (2004) suggests that the potential threat of outliers shrinks with sample size so that the potential influence of low/high probabilities, providing $1/\left[N \times (1-\hat{p}_{max})\right]$ and $1/\left[N \times (\hat{p}_{min})\right]$ as bounds for the influence of observations at the top and bottom of the probability distribution. Using the sample that lies along the range of the treatment observations ($[0.001, 0.811]$) limits the potential effect of outlier treatment and control impact on the estimator to less than 2.5%, well within the bounds used in other studies.

More importantly in the context of Rose (2017), when truncating the sample in this way, the difference in means across observable characteristics of EMU and non-EMU groups falls much closer to zero, further still when weighted by inverse propensity scores. It would appear in large trade datasets that the *opposite* problem to that of influential outliers may be problematic for the Equation 3. Rather than being driven by small(large) treatment(control) outliers, estimates on the full sample are strongly influenced by the large number of Non-EMU observations with extremely low probability of treatment. These receive smaller weights in Equation 3, but the weighting penalty only slightly diminishes this group's out-sized effect on the estimators, as I will show.This problem is larger in unweighted estimates of Equation 1 and Equation 2. Another way this can be shown is by *censoring* p-scores (keeping the observation but setting an upper/lower threshold for their weights), which has little effect relative to IPWRA estimates on the full data. The fact that truncating these observations changes their results dramatically, while assigning arbitrary lower/upper probability limits on the full sample does not, implies that the large weight of low probability observations are dramatically altering estimates in spite of their low weights, not because of them. For clarity of exposition I do not include such censored results below.

I agree with Rose (2017) that sample choice is critical in the estimation of EMU trade effects. Before presenting estimates of the EMU I show that *how* the sample is selected points towards preferring estimates made on a smaller subset of data. I consider five samples: the full sample including zero and missing trade flows,[7] the full sample with non-zero trade (ie using the log specification), that for which predicted probabilities of treatment are non-zero (ie the largest sample where the IPWRA is estimable), a truncated sample with probabilities limited to the range of the treatment group, and finally an *ad hoc* sample of upper income countries. This last group is consistent with the definition in Rose

[7]For PPML cases I assume missing trade data reflects a zero trade flow in cases where the GDP data for origin and destination countries is non-missing, leaving remaining cases as missing.

Figure 2: *Difference in Means (Treatment - Control), Various Control Samples*



(2017), and includes only origin and destination countries with real GDP per capita greater than $12,736.

Figure 2 shows the difference in means between EMU (treatment) and non-EMU (control) observations across each of these samples for each of the controls used in Equation 4. I provide both the unweighted and weighted differences in means using propensity score estimators when possible. In an ideal situation, such as a properly run randomized control trial, these groups should be identical along observable characteristics and these differences should be zero. This exercise highlights the value that the first stage propensity score process has in improving the comparability of these two groups.

The absolute difference in means for the full sample is systematically largest, with the non-missing trade already substantially improving the comparability between the two groups. However in the full log sample, which is quite close to that used in Glick and Rose (2016), differences are still extremely far from zero. Because these differences are precisely estimated, many of the error bands are not visible, but they are strongly significant. The improvement when moving from full to non-zero/missing trade is somewhat intuitive given that these flows are concentrated in small economies. Comparing the largest p-score samples, weighting improves comparability, but in some cases only marginally. Truncating the sample to include only the non-EMU observations that fall within the range of estimated p-scores of EMU pairs brings these much closer to zero, with weighting closing the remaining gap substantially for all but one (difference in GDP per capita) control. Though these small differences are in some cases still significant they provide a much more sensible experimental control group than the full sample estimates.

It is interesting to compare these model driven sample means with the ad hoc, "upper income" sample used in Rose (2017). For the control used to restrict the sample (GDP per

capita) this actually outperforms the comparability made using propensity scores. The model in Equation 4 works to minimize gaps along all of these variables, many of which substantially outperform the ad hoc measure, at the expense of fitting the per-capita GDP comparisons as closely. The final sample, which combines the ad hoc weights with the logistic model propensity scores does improve the fit (though not generally as well as the p-score trucation), but I caution that this actually limits the sample as a non-trivial share of the upper-income countries were allocated scores of zero and thus drop out when the inverse-propensity score weights are included. I take the strength of these ad hoc measures as evidence that they, or a mixed approach, may prove fruitful in situations where the researcher has strong reasons to limit the sample along margins that are difficult to econometrically model.

One can consider a reduction in sample using IPWRA estimates as a data-driven approach to the sample selection, providing the best possible comparability between treated and control, conditional on observable characteristics, while also improving comparability through the weighting procedure itself. Of course, such attempts at constructing a re-randomized treatment and control set ex-post are only as reliable as the observable characteristics used in the first-stage selection model. It is still possible that unobserved characteristics not captured in this process may bias estimates. I note that this only creates a problem if they are *also* not captured by the rich set of origin-year, destination-year, and dyadic fixed effects used in the gravity equation. While this is still potentially true, my results should improve upon the existing estimates of the EMU treatment effect in the aggregate trade data by providing a nearly balanced treatment and control group prior to estimating the marginal effect using conventional gravity measures.

4. Results: IPWRA Gravity Estimates of Trade

I now present results for Equation 1 and Equation 2 on the samples described above, showing when possible the improvements made by weighting as in Equation 3. Because I wish to understand which changes come from weighting and which come from sample selection, I will include the OLS/PPML without the IPWRA estimate for each of these sub-samples. I begin with the OLS estimates of the log gravity specification in Table 2.

Table 2: *Log Gravity Estimates*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EMU	0.483^{***}	0.477^{***}	0.409^{***}	0.069^{***}	0.057^{***}	0.133^{***}	0.024
	(0.017)	(0.017)	(0.017)	(0.022)	(0.021)	(0.028)	(0.026)
Regional Trade Agreement	0.323^{***}	0.248^{***}	0.246^{***}	0.025	0.043^{*}	0.059^{*}	$0.053^{\dagger}$
	(0.008)	(0.009)	(0.009)	(0.026)	(0.025)	(0.030)	(0.032)
Exporter-Year FEs	✓	✓	✓	✓	✓	✓	✓
Impoter-Year FEs	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓
Sample: p-score Non-missing		✓	✓				
Sample: p-score Truncated				✓	✓		
Sample: ad hoc, upper income						✓	✓
IPWRA			✓		✓		✓
R2	0.871	0.880	0.888	0.967	0.974	0.949	0.964
N	810018	603870	603870	40447	40447	75311	47727

All estimates use OLS estimates with log bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis, $\dagger p < 0.15, *p < 0.10, **p < 0.05, ***p < 0.01$.

The estimate using the full sample is slightly larger, at 0.48 than the 0.43 coefficient from Glick and Rose (2016). My starting point is therefore an EMU effect that implies a 61% ($e^{0.48} - 1$) increase in bilateral trade as a result of the euro. This effect is essentially unchanged when running the same specification on the smaller sample that omits propensity scores assigned an exact zero by the first stage model. This is perhaps not too surprising given that these samples have fairly similar differences in means across treatment and control in Figure 2. When applying the IPWRA estimator to weight the regression outcomes by probability of selection this estimate is substantially reduced to 0.41, but remains large and quite close to the Glick and Rose (2016) estimate for the EMU. Estimates on the smaller, truncated, sample using only observations within the range of the propensity scores observed for the EMU, (in the range $[0.001, 0.811]$) decrease these effects substantially to a roughly 6-7% increase in trade. Again the weighted effect is slightly smaller, but is now not statistically distinguishable from the unweighted. The propensity score is still important for the unweighted sample given that it was used to select inclusion, but this selection matters much more than weighting. For the ad hoc Upper income sample, I estimate a 0.13 EMU effect, quite similar to the equivalent estimate of 0.11 in Rose (2017). Estimates using this sample selection are nearly twice as large as those on the truncated propensity score sample. Adding the propensity score weights to this sample to estimate the IPWRA treatment effect completely reduces this effect to zero, though notably there

is a large fraction of the this sub-sample that have zero estimated propensity scores, and therefore drop out of this estimation. Repeating the estimation using the sample in column 7 without weighting provides an estimate of 0.03, so as with the truncated sample most of this difference appears to come from the selection effect of dropping observations with extremely poor fit in the first stage model, rather than the weighting estimator itself.

Repeating this exercise for the PPML estimates of the gravity equation provide substantially different results, and are shown in Table 3. The first column uses the full sample, assuming zero trade flows for missing values where the macroeconomic controls (GDP and GDP per capita) are non-missing. The resulting estimate of the EMU effect is 0.085, larger than the equivalent estimate of 0.052 in Larch et al. (2019), which again I attribute to using a different sample and not jointly controlling for other currency unions (rather dropping their observations). Now when reducing the sample to those with positive propensity scores this falls only slightly to 0.06, and again to 0.04 with the truncated propensity score sample. These estimates for the PPML are all quite close to those from the log-gravity estimation using the truncated sample.

Table 3: *PPML Gravity Estimates*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EMU	0.085***	0.059***	0.055***	0.045**	0.041*	0.006	-0.002
	(0.011)	(0.011)	(0.011)	(0.023)	(0.023)	(0.014)	(0.015)
Regional Trade Agreement	0.057***	0.128***	0.129***	0.079**	0.082***	0.070***	0.066***
	(0.010)	(0.007)	(0.007)	(0.031)	(0.032)	(0.014)	(0.015)
Exporter-Year FEs	✓	✓	✓	✓	✓	✓	✓
Impoter-Year FEs	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓
Sample: p-score Non-missing		✓	✓				
Sample: p-score Truncated				✓	✓		
Sample: ad hoc, upper income						✓	✓
IPWRA			✓		✓		✓
pseudo-R2	0.991	0.991	0.992	0.996	0.997	0.996	0.996
N	1521887	603870	603870	40447	40447	86971	47727

All estimates use Poisson pseudo-maximum likelihood estimators with bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis, $\dagger p < 0.15, *p < 0.10, **p < 0.05, ***p < 0.01$.

While my preferred truncated propensity score sample estimates remain robust in this PPML estimation using heteroskedastic-robust standard errors, Egger and Tarlea (2015) suggest using multi-way exporter, importer, and year clustering in gravity equation

estimations. Consistent with the PPML results of Larch et al. (2019), who cite both Huber/White and multi-way clustered errors of this type, the significance for all of my results using the PPML estimator fail to reach even a $p < 0.15$ threshold when implementing these standard errors. I report these more forgiving standard errors here as they are those used in Glick and Rose (2016) and Rose (2017), with which I wish to draw closest comparison. I caution that in addition to all PPML estimates, the log gravity estimates on the truncated sample also lose statistical significance using exporter-importer-year multi-way clustering.[8]

An interesting result in Table 3 is the consistency of estimates across samples in columns 2-5. Once dropping the observations with an exact zero estimated probability of treatment the PPML estimator appears quite robust to sample selection, with all estimates extremely close to my preferred specification using the log-gravity specification in column 5 of Table 2. It is also notable that the ad hoc sample, which in general was a poorer fit than the truncated p-score sample in Figure 2, has lower estimates. I interpret the results in this table as suggesting that the PPML estimate is considerably more robust to the selection issues from log-gravity specifications. While the full and ad-hoc samples differ dramatically, the large reduction in sample when using the full propensity score sample to the truncated sample results in only small differences in my estimates. The differences when using standard PPML and the IPWRA framework are also small.

5. Conclusions

Eurozone membership was not assigned by a researcher, but the culmination of intense policy deliberation. While it may not be possible to fully rectify selection effects of this policy's effect on trade, propensity score methods offer a way of improving the credibility of estimators to take such differences between EMU and non-EMU pairs into account. Log gravity equation estimates of the EMU treatment effect are extremely sensitive to sample changes. I argue that this sensitivity, as reported in the existing literature, largely reflects improvements in the control sample, bringing their observable characteristics closer to those of EMU countries. Ad hoc measures do a good job fitting on a given statistic, as my example using a GDP per capita cutoff to determine selection into the estimation sample, but I argue that data driven propensity score approaches provide a better way of improving the comparability of these groups, and does not rely on arbitrary decision making by researchers. Use of the doubly robust IPWRA estimator provides a method for re-weighting the average treatment effects from gravity equations to further improve this fit, but in practice has only small quantitative implications relative to the effects of truncation.

[8]Though the results in columns 1-4 of Table 2 remain significant (at varying levels) when using more conventional exporter or pairwise clustering.

Results using samples where observable EMU and non-EMU characteristics are most comparable suggest a small and positive impact of the currency union on trade. This is roughly a 4-5% increase that is quite similar across my preferred log and PPML specifications. These are in line with the survey of the literature by Polák (2019), and explain why the large estimates of Glick and Rose (2016) are likely biased upwards. While Rose (2017) identifies the correct reason for differences in trade estimates of the euro, the evidence presented here suggests that his conclusion that more data should be better, is likely mistaken.

More broadly my results suggest that greater care should be made in sample selection in trade and other macro-policy environments. Observational studies can leverage models of first stage treatment to ensure better balance between countries that are exposed to a particular policy, and those that are not. While this does not completely bridge the gap in causal identification between observational studies and the experimental ideal, it provides an empirically driven step in the right direction. The PPML estimator is much more robust to changing sample, though still reflects statistically meaningful shifts when moving from the full sample to those that drop the most extreme outliers. This suggests that research using log approximations should be particularly careful of these sample selection issues.

References

Baldwin, R. and D. Taglioni (2007). Trade effects of the euro: A comparison of estimators. *Journal of economic integration*, 780–818.

Baldwin, R. E. (2006). The euro's trade effects.

Campbell, D. L. (2013). Estimating the impact of currency unions on trade: solving the glick and rose puzzle. *The World Economy 36*(10), 1278–1293.

Chintrakarn, P. (2008). Estimating the euro effects on trade with propensity score matching. *Review of International Economics 16*(1), 186–198.

Conte, M., P. Cotterlaz, and T. Mayer (2021). The cepii gravity database. *CEPII: Paris, France*.

Egger, P. H. and F. Tarlea (2015). Multi-way clustering estimation of standard errors in gravity models. *Economics Letters 134*, 144–147.

Glick, R. and A. K. Rose (2002). Does a currency union affect trade? the time-series evidence. *European economic review 46*(6), 1125–1151.

Glick, R. and A. K. Rose (2016). Currency unions and trade: A post-emu reassessment. *European Economic Review 87*, 78–91.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of international economics*, Volume 4, pp. 131–195. Elsevier.

Hirano, K. and G. W. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology 2*(3), 259–278.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics 86*(1), 4–29.

Jordà, Ò. and A. M. Taylor (2016). The time for austerity: estimating the average treatment effect of fiscal policy. *The Economic Journal 126*(590), 219–255.

Kenen, P. B. (2002). Currency unions and trade: Variations on themes by rose and persson. *Reserve Bank of New Zealand Discussion Paper No. DP2002/08*.

Larch, M., J. Wanner, Y. V. Yotov, and T. Zylkin (2019). Currency unions and trade: A ppml re-assessment with high-dimensional fixed effects. *Oxford Bulletin of Economics and Statistics 81*(3), 487–510.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine 23*(19), 2937–2960.

Micco, A., E. Stein, and G. Ordoñez (2003). The currency union effect on trade: early evidence from emu. *Economic policy 18*(37), 315–356.

Millimet, D. L. and R. Tchernis (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business & Economic Statistics 27*(3), 397–415.

Nitsch, V. (2002). Honey, i shrunk the currency union effect on trade. *The World Economy 25*(4), 457–457.

O'Rourke, K. H. and A. M. Taylor (2013). Cross of euros. *The Journal of Economic Perspectives*, 167–191.

Persson, T. (2001). Currency unions and trade: how large is the treatment effect? *Economic policy 16*(33), 434–448.

Polák, P. (2019). The euro's trade effect: a meta-analysis. *Journal of Economic Surveys 33*(1), 101–124.

Rose, A. (2002). Honey, the currency union effect on trade hasn't blown up. *The World Economy 25*(4), 475–479.

Rose, A. K. (2000). One money, one market: the effect of common currencies on trade. *Economic policy 15*(30), 08–45.

Rose, A. K. (2001). Currency unions and trade: the effect is large. *Economic Policy 16*(33), 449–461.

Rose, A. K. (2017). Why do estimates of the emu effect on trade vary so much? *Open Economies Review 28*(1), 1–18.

Rose, A. K. and P. Honohan (2001). Currency unions and trade: the effect is large. *Economic Policy*, 449–461.

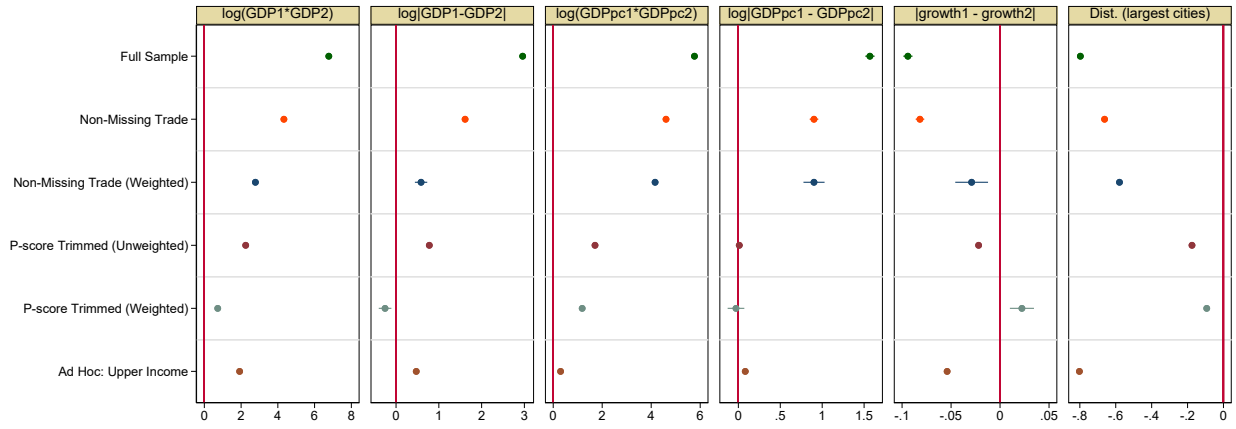
Silva, J. S. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641–658.

Silva, J. S. and S. Tenreyro (2011). Further simulation evidence on the performance of the poisson pseudo-maximum likelihood estimator. *Economics Letters 112*(2), 220–222.

Słoczyński, T. and J. M. Wooldridge (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory 34*(1), 112–133.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics 141*(2), 1281–1301.

Figure 3: *Difference in Means (Treatment - Control), Various Control Samples: simplified logistic*



A. Estimates using an alternative first stage specification

My preferred specification uses the linear and squared terms of regression controls in Equation 4. The case could be made for the simpler linear model that excludes these squared terms, given that the fit is only marginally improved and that fewer observations are deleted from both the non-zero p-score and truncated sub-samples. In this section I show that doing so has little bearing on my results, yielding similarly strong, though slightly weaker convergence of treatment and control means in the truncated sample, and nearly identical regression results overall. First, using this simplified first stage I replicate Figure 2 in Figure 3.

For the absolute difference of per capita GDP this model now generates closer means across EMU and non-EMU countries in the truncated sample, but otherwise has slightly poorer comparability. Both log-OLS and PPML regression results are nearly the same, as is clear from Table 4 and Table 5. For both estimators the truncated IPWRA estimator is identical for the PPML estimator, and slightly larger for the log-gravity OLS estimation (0.08, instead of 0.06). Given that these are nearly it's likely the case that the information gain from including squared terms in the main body of the paper is only marginal, and all of the results discussed above hold when using a simplified model that includes more data.

Table 4: *Log Gravity Estimates*

	(1)	(2)	(3)	(4)	(5)	(6)
EMU	0.48^{***}	0.50^{***}	0.42^{***}	0.10^{***}	0.08^{***}	0.13^{***}
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Regional Trade Agreement	0.32^{***}	0.30^{***}	0.30^{***}	0.12^{***}	0.12^{***}	0.06^{**}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Exporter-Year FEs	✓	✓	✓	✓	✓	✓
Impoter-Year FEs	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓
Sample: p-score Non-missing		✓	✓			
Sample: p-score Truncated				✓	✓	
Sample: ad hoc, upper income						✓
IPWRA			✓		✓	
R2	0.871	0.877	0.888	0.949	0.957	0.949
N	811337	712609	712609	95032	95032	75456

All estimates use OLS estimates with log bilateral export flows as the dependent variable. Standard errors clustered by country-pairs, $\dagger p < 0.15, *p < 0.10, **p < 0.05, ***p < 0.01$.

Table 5: *PPML Gravity Estimates*

	(1)	(2)	(3)	(4)	(5)	(6)
EMU	0.09^{***}	0.06^{***}	0.07^{***}	0.02	0.04^{**}	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Regional Trade Agreement	0.06^{***}	0.07^{***}	0.07^{***}	0.14^{***}	0.14^{***}	0.07^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Exporter-Year FEs	✓	✓	✓	✓	✓	✓
Impoter-Year FEs	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓
Sample: p-score Non-missing		✓	✓			
Sample: p-score Truncated				✓	✓	
Sample: ad hoc, upper income						✓
IPWRA			✓		✓	
pseudo-R2	0.991	0.991	0.992	0.995	0.996	0.996
N	1523588	712609	712609	95032	95032	87117

All estimates use Poisson pseudo-maximum likelihood estimators with bilateral export flows as the dependent variable. Standard errors clustered by country-pairs, $\dagger p < 0.15, *p < 0.10, **p < 0.05, ***p < 0.01$.