The WTO Trade Effect*

Pao-Li Chang†
School of Economics
Singapore Management University

Myoung-Jae Lee‡
Department of Economics
Korea University
and
Research School of Economics
Australian National University

May 20, 2011

Abstract

This paper re-examines the GATT/WTO membership effect on bilateral trade flows, using nonparametric methods including pair-matching, permutation tests, and a Rosenbaum (2002) sensitivity analysis. Together, these methods provide an estimation framework that is robust to misspecification bias, allows general forms of heterogeneous membership effects, and addresses potential hidden selection bias. This is in contrast to most conventional parametric studies on this issue. Our results suggest large GATT/WTO trade-promoting effects that are robust to various restricted matching criteria, alternative GATT/WTO indicators, non-random incidence of positive trade flows, inclusion of multilateral resistance terms, and different matching methodologies.

JEL Classification: F13; F14; C14; C21; C23

KEY WORDS: Trade flow; Treatment effect; Matching; Permutation test; Signed-rank test; Sensitivity analysis.

---

*We thank two anonymous referees, Robert Staiger, Jeffrey Bergstrand, and Alan Deardorff for helpful comments and suggestions. We also thank colleagues, Hian Teck Hoon, Davin Chor, and Tomoki Fujii, at SMU for helpful feedback at different stages of the project. The research of Myoung-jae Lee was supported by the Korea Research Foundation funded by the Korean Government (KRF-2009-327-B00091).

†School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903. Email: plchang@smu.edu.sg. Tel.: +65-68280830. Fax: +65-68280833.

‡Department of Economics, Korea University, Anam-dong, Sungbuk-gu, Seoul 136-701, South Korea. Email: myoungjae@korea.ac.kr. Tel./Fax: +82-2-32902229. Research School of Economics, College of Business and Economics, Australian National University, Canberra ACT 0200, Australia.

1. INTRODUCTION

Since its creation in 1947, the General Agreement on Tariffs and Trade (GATT) has played an important role in the international trading system. It has sponsored eight rounds of trade-policy negotiations that successfully brought down the average tariff rates on industrial goods and also expanded the set of substantive rules governing international trade (beyond tariffs to nontariff barriers, and beyond trade in merchandise to trade in services). This process culminated in the establishment of the World Trade Organization (WTO) in 1995. Since 1947, the GATT/WTO has also grown in its membership from a small set of 23 (mainly developed) countries to a roster that now includes more than 150 countries. Meanwhile, global trade flows have increased exponentially at a rate above the growth rate of merchandise output. It is against this backdrop that the finding by Rose (2004) came as a surprise.

Based on the gravity model of trade (that hypothesizes that the bilateral trade volume between two countries varies positively with their economic sizes and inversely with their bilateral trade resistance), Rose (2004) conducted parametric estimations and found that the GATT/WTO membership status of a country pair had no statistically significant effect on bilateral trade. This negative finding was partially reversed by Tomz et al. (2007) when they reclassified countries according to their participation status in the GATT/WTO (instead of formal membership), and by Subramanian and Wei (2007) when they differentiated the effects by subsets of the sample (e.g., developed versus developing countries). Although shedding light on possible caveats to the original study by Rose (2004), these studies and other follow-up research in this literature have largely followed the conventional approach of parametric estimation. In this paper, we argue that when the leading gravity theories do not have clear guidance on the parametric (functional) relations of the empirical trade-resistance measures, and when the economic theories of trade agreements (e.g., Bagwell and Staiger, 2010, pp. 245–247) suggest that heterogeneous membership effects on trade are important implications (of uneven levels of trade negotiation participation), these existing parametric studies are at risk of misspecification bias on both accounts. We propose a system of nonparametric methods that is geared toward these concerns to re-evaluate the GATT/WTO trade effect.

In particular, we apply pair-matching methods to obtain point effect estimates. Following the established gravity theories (Anderson, 1979; Bergstrand, 1985; Deardorff, 1998; Anderson and van Wincoop, 2003), empirical researchers have come to adopt a long list of variables as proxies for the theoretical concept of trade resistance between a pair of countries. This list typically includes (foremost) distance, geographic characteristics, language, colonial ties, currency union, free trade agreement, and the GATT/WTO membership status. However, there is no clear theoretical justification for the linear relation (among the various trade-resistance measures) that is often adopted in the empirical studies. In this paper, we conduct matching based on a set of covariates that is exactly the same as the list of regressors used in parametric studies. However, by matching observations that have different treatment status but are otherwise similar in terms of these covariates, we do not have to take a stand on the functional relations among these observed covariates and hence avoid potential parametric misspecifications. In addition, the matching method by design allows for the treatment (i.e., membership) effect to vary with the observed covariates, and thus it can accommodate arbitrary forms of heterogeneous treatment effects. In general, the homogeneous effect estimate in regression approaches does not correspond to the average of subject-wise heterogeneous effects, if the heterogeneity takes on highly nonlinear functional forms.

We also address other potential econometric concerns arising in the current application. First, given a panel of bilateral trade data, which likely have a complicated data structure with serial and spatial dependence, this paper applies permutation tests that circumvent the difficulty in deriving asymptotic tests. Permutation tests are nonparametric and exact inferences (applicable to finite sample sizes). They are also straightforward to implement in the matching framework. We generalize the test to explicitly allow for heterogeneous treatment effects in constructing the confidence intervals. Finally, we complete the estimation procedure with a nonparametric sensitivity analysis à la Rosenbaum (2002) to formally address potential bias due to unobserved self-selection into membership. We put together the above methods in a coherent manner such that they can be easily applied to other treatment effect problems of a similar nature.

Applying the nonparametric methods to the data set of Rose (2004), we reach a conclusion that is in stark contrast with Rose (2004): membership in the GATT/WTO has large and significant trade-promoting effects. We explore robustness of this result to various possible caveats; the general finding continues to hold. First, both parametric gravity and nonparametric matching estimators rely on the assumption of 'selection on observables'; in other words, non-random selection into membership based on unobservables is assumed away. This assumption may fail if there are important omitted variables. The Rosenbaum (2002) sensitivity analysis partly addresses this problem. Alternatively, we also conduct restricted matching, where we further limit the match to observations from the same 'dyad' (where a dyad indicates a pair of trading countries), the same year, or the same relative development stage. This eliminates potential bias arising from unobserved heterogeneity across dyads, years, or development stages.

Second, Tomz et al. (2007) emphasize the importance of *de facto* participation in the GATT/WTO by colonies, newly independent nations and provisional members, and find strong GATT/WTO effects on trade when this type of nonmember participation is taken into account. We conduct the same nonparametric analysis using the data set of Tomz et al. (2007) and find even stronger results than those based on the Rose (2004) data set.

Third, we verify the robustness of pair-matching by conducting 'kernel-weighting matching', which allows multiple matches for a subject while assigning greater weights to closer matches. The kernel-weighting matching effect estimates are very similar to pair-matching estimates.

Fourth, by using the data set of Rose (2004) or Tomz et al. (2007), we have based our analysis on observations with positive trade flows. Studies by Helpman et al. (2008) and Felbermayr and Kohler (2007) suggest that the incidence of positive trade flows may not be random. To address possible bias due to non-random incidence of active trading relationships, we apply our nonparametric procedures to the subset of the data where a dyad has reported bilateral trade flows before either country in the dyad ever joins the GATT/WTO. For these observations, the membership effect on prompting new trading relationships is not relevant, and hence the effect estimates correspond to only the membership effect on trade volumes. We find overall stronger effect estimates based on this refined analysis.

Fifth, relative, rather than absolute, trade resistance is argued by some gravity theories to be more appropriate in explaining bilateral trade flows (Anderson and van Wincoop, 2003); thus, multilateral resistance terms may have to be controlled for. We follow recent studies by Baier and Bergstrand (2009a,b) to approximate the endogenous multilateral resistance terms by observable exogenous trade resistance covariates in the matching framework. The strong trade effects of GATT/WTO remain.

Finally, we explore an alternative treatment effect concept, difference-in-difference, which is based on weaker identification assumptions and thus could be more robust to potential bias due to selection on unobservables. This method compares the difference over time in the trade volume of a member dyad to that of a comparable nonmember dyad. The matching estimates indicate that the GATT/WTO trade effects are negligible in early phases of the membership, but become statistically and economically significant five or six years after the GATT/WTO accession. To complete the analysis, we conduct placebo exercises and verify that the time trends of trade flows of matched dyads are the same in advance of membership, dismissing concerns that the difference-in-difference estimates may be picking up systematic differences in time trends between member and nonmember dyads due to unobservables not controlled for.

The discrepancy between the finding of the current nonparametric approach and that of the conventional parametric approach suggests that parametric gravity models may be misspecified. We explore generalizing the parametric gravity model's specifications to reduce the discrepancy. Our limited search suggests that the assumption of homogeneous membership effects could be a major source of misspecification. By allowing the membership dummies to interact with observed covariates (and hence allowing the membership effects to vary with dyad-year characteristics), we find the parametric effect estimates to become significant and positive. However, more research into the nature of heterogeneous membership effects seems desirable and we leave this for future research.

The rest of the paper is organized as follows. Section 2 introduces the nonparametric methodologies. Section 3 explains the data used. Sections 4 and 5 present our benchmark estimation results and robustness checks. Section 6 explores potential misspecifications of the parametric gravity models. Section 7 provides our conclusions.

2. METHODOLOGY

2.1 Mean Effects and Matching

Recall that a 'dyad' indicates a pair of trading countries. In the current application, an observation unit corresponds to a dyad i in a year t, while a matched 'pair' indicates two observation units matched on covariates. Let d_{it} denote the observed treatment status of a dyad i in year t, where $d_{it} = 1$ if the subject it is treated and 0 if untreated. The treatment dummy d_{it} takes on different meanings as the treatment under study changes. For example, a dyad-year is 'both-in' treated if both countries of the dyad in the year are GATT/WTO members and untreated if both are nonmembers. Define y_{it}^1 (y_{it}^0) as the potential *treated* (*untreated*) response; in our application, this corresponds to the potential treated (untreated) bilateral trade volume of the dyad-year it. Thus, the observed response is $y_{it} \equiv d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0$. Finally, let x_{it} denote the observed covariates for the dyad-year that could potentially affect the selection into the treatment and the response to the treatment. Label the group of treated and untreated observations 'the treatment group' and 'the control group', respectively. In the following, we will often omit the subscript it to simplify presentation.

One can identify the mean effect $E(y^1 - y^0|x)$ conditional on x by the conditional group mean difference:

$$E(y|d = 1, x) - E(y|d = 0, x) = E(y^1|d = 1, x) - E(y^0|d = 0, x) = E(y^1 - y^0|x) \quad \text{if } (y^0, y^1) \amalg d|x,$$

where $(y^0, y^1) \amalg d|x$ is the identifying 'selection on observables' assumption. It says that both the potential treated and untreated responses (y^0, y^1) are independent of d given x; that is, the only source of selection bias is via the observed covariates; the selection into treatment is random once x is controlled for. This identifying condition is actually equivalent to the condition in parametric regression approaches that the treatment dummy be uncorrelated with the error term of the regression.

A weaker identifying assumption $y^0 \amalg d|x$ is sufficient if one is only interested in the 'effect on the treated', as under the assumption,

$$\begin{aligned} E(y|d = 1, x) - E(y|d = 0, x) &= E(y^1|d = 1, x) - E(y^0|d = 0, x) \\ &= E(y^1|d = 1, x) - E(y^0|d = 1, x) = E(y^1 - y^0|d = 1, x). \end{aligned}$$

Alternatively, the assumption $y^1 \amalg d|x$ is sufficient to identify the 'effect on the untreated' $E(y^1 - y^0|d = 0, x)$. Once the x-conditional effect is found, x can be integrated out to yield a marginal effect. For example, for the effect on the treated, the distribution $F(x|d = 1)$ of $x|d = 1$ can be used to obtain the mean effect:

$$E(y^1 - y^0|d = 1) = \int E(y^1 - y^0|d = 1, x)\mathrm{d}F(x|d = 1).$$

This framework of first finding the x-conditional effect has two obvious advantages: first, by conditioning on x, we do not need to model the structural relationship among x and avoid the misspecification bias that may arise in the parametric approach; second, this allows for trade effects to vary with dyad-year characteristics x in arbitrary ways. The unconditional mean effect then reflects the average of the heterogeneous x-conditional treatment effects weighted by the frequency of x. It is this average effect (on all, on the treated, or on the untreated) that we estimate. This departs from the parametric gravity regression approach, where a homogeneous treatment effect regardless of x is typically assumed. Apparently, by conditioning on x, self-selection into treatment based on the observed covariates is controlled for in the matching framework.

The pair-matching estimator for the effect on the treated can be obtained as follows. First, consider a treated subject, say subject it. Second, select the control subject that is the closest to the treated subject it in terms of x.[1] Third, suppose M matched pairs are obtained, and y_{m1} and y_{m2} are the trade volumes of the two subjects in pair m ordered such that $y_{m1} > y_{m2}$ without loss of generality. Then, defining $s_m = 1$ if the first subject in pair m is treated and -1 otherwise, the effect on the treated can be estimated with

$$D \equiv \frac{1}{M} \sum_{m=1}^{M} s_m (y_{m1} - y_{m2}) \rightarrow^p E(y^1 - y^0 | d = 1) \quad \text{under } y^0 \amalg d | x, \tag{1}$$

which is simply the average of the pair-wise differences in trade volumes of treated and untreated subjects in matched pairs.

Some remarks are in order. First, for a treated subject, if there is no good matching control, the subject may be passed over; i.e., a 'caliper' c may be set such that a treated subject it with $\min_{i't' \in C} \|x_{it} - x_{i't'}\| > c$ is discarded, where '$i't' \in C$' indicates subjects in the control group. Second, the above matching scheme can be reversed to result in an estimator for the effect on the untreated: consider the control group, and for each control subject, select the best matching subject from the treatment group. Finally, one can estimate the effect on all by including, in the estimator D, all (treated and control) subjects that have a qualified match.

Matching is widely used in labor and health economics. See, for example, Heckman et al. (1997) and Imbens (2004), and applications in Heckman et al. (1998), Lechner (2000), and Lu et al. (2001). Matching methods have also started to appear in international economics studies such as Persson (2001) on the currency union effect and Baier and Bergstrand (2009b) on the free trade agreement effect. See Rosenbaum (2002) and Lee (2005) for more discussions on treatment effects and matching in general.

[1] We use the simple scale-normalized distance measure, $(x_{it} - x_{i't'})\Sigma_x^{-1}(x_{it} - x_{i't'})'$, where $i't'$ refers to a control subject and Σ_x is a diagonal matrix containing the sample variances of the covariates in the pooled sample on the diagonal. As x in our data includes continuous variables (cf. Section 3), the likelihood of multiple-matching (multiple control subjects with the same distance to the treated subject) is negligible; thus, we restrict our attention to pair-matching (where each subject has a unique closest match).

2.2 Permutation Test for Matched Pairs

Although matching estimators are popular in practice, their asymptotic properties are not fully understood.[2] In practice, a standard t-statistic or a bootstrap procedure is often used to derive the p-value or the confidence interval (CI). The standard t-statistic is straightforward but theoretical justifications are not available in most cases; on the other hand, the bootstrap is computationally demanding and argued to be invalid by Abadie and Imbens (2006). In this paper, we propose using permutation tests.

Permutation tests invoke the concept of *exchangeability*, which suggests that under the null hypothesis H_0 of no effect, potential treated and untreated responses are exchangeable without affecting their joint distribution: $F(y_{it}^0, y_{it}^1|x) = F(y_{it}^1, y_{it}^0|x)$. This implies that under the null, the two potential responses have the same marginal distribution and hence the same mean given x. Thus, we can test the equal mean (i.e., zero mean effect) implication of the null.

It is straightforward to carry out the permutation test described above for matched pairs and test for a zero mean effect under the null. Under the null hypothesis of exchangeability, the two subjects in each matched pair are exchangeable in the labeling of their treatment status (treated or untreated). In each permutation of 'pseudo' treatment assignment, one can calculate the 'pseudo' effect estimate. By obtaining all possible 2^M permutations of the treatment labels in all M pairs, one can calculate the exact p-value of the observed mean effect estimate D by placing it in the 'empirical' distribution of the pseudo effect estimates.

When M is large (as in the current application), such that the number of permutations is huge, one can approximate the exact p-value by simulating only a subset (say, 1000) of permutation possibilities from the complete permutation space and comparing the observed effect estimate D against the simulated sample of pseudo effect estimates. Alternatively, one can apply normal approximation. Note that in a permutation, the obtained pseudo effect estimator can be written as $D' \equiv \frac{1}{M}\sum_{m=1}^{M} w_m s_m(y_{m1} - y_{m2})$, where w_m, $m = 1, ..., M$, is a *iid* random variable such that $P(w_m = 1) = P(w_m = -1) = 0.5$. That is, the treatment labels of the two responses in pair m are exchanged if $w_m = -1$, and no exchange otherwise. We show in the appendix that, conditional on the observed data, the exact p-value of D can be approximated by

$$P(D' \geq D) \simeq P\left\{N(0,1) \geq \frac{D}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}}\right\}, \tag{2}$$

which turns out to use the same t-statistic as the conventional two-sample test. Thus, this display incidentally provides a theoretical justification for the common practice of using the t-statistic to evaluate the significance of matching estimators, although we have derived (2) from an exact inferential approach (i.e., permutation with respect to the treatment labels but conditional on the observed data) and not based on asymptotic distribution theories (i.e., sampling with respect to the data).

[2] See, however, exceptions such as Abadie and Imbens (2006) for the case of *iid* data.

In addition to testing the null hypothesis of a zero mean effect, one may also be interested in an interval estimate of the mean effect. We show in the appendix how to obtain the CI for the mean effect by 'inverting' the above test (e.g., Lehmann and Romano, 2005). It is worth noting that in deriving the CI, we have generalized the inverting procedure to explicitly allow for heterogeneous treatment effects.

As indicated above, permutation inference methods have several advantages: (i) they are non-parametric as they do not require distributional assumptions on the response, other than the exchangeability condition, and (ii) they are exact inferences despite making no parametric distributional assumptions in small samples, and they are often equivalent to conventional asymptotic inference methods in large samples when normal approximation is used. On the other hand, as permutation tests invoke a stronger concept of no effect (on the distribution), this rules out testing for null hypotheses of no effect (on the mean) that still allow some effects on other moments of the distribution. In small samples where normal approximation does not apply, permutation tests may also be computer-intensive. Both disadvantages, however, are not important in the current application.

Permutation tests, instead of asymptotic tests, are especially convenient in the current application with a panel of bilateral trade data, which possibly have a complicated data structure with serial and spatial dependence, rendering the derivation of asymptotic properties for the matching estimator difficult if not impossible. By relying on exchangeability as the null hypothesis of no effect, the permutation test can accommodate potentially a wide range of data structures. For example, suppose that the joint distribution $F(y_{it}^0, y_{it}^1|x)$ is normal. In this scenario, the exchangeability condition requires only that the treated and untreated responses have the same mean and variance conditional on x. This allows for heteroskedasticity (i.e., variances of responses to vary with x) or correlation across time or observation units.

Permutation tests have a long history in statistics since Fisher (1935) and are widely used in statistics and medicine. Recently, Imbens and Rosenbaum (2005) applied permutation inference to well-known "weak instrument" data in economics to find that only permutation methods provided reliable inference. Ho and Imai (2006) also applied permutation inference to a political science data set. As can be seen in these examples, the application of permutation methods is fairly new in the social sciences. For more on permutation (or randomization) tests in general, see Hollander and Wolfe (1999), Pesarin (2001), Ernst (2004), and Lehmann and Romano (2005), among others.

2.3 Signed-Rank Test for Matched Pairs

Instead of the difference in response $s_m(y_{m1} - y_{m2})$, we can apply the permutation inference to the 'signed rank' of the difference in response. The advantage is that rank-based tests are more robust to outliers. In addition, the ensuing Rosenbaum (2002) sensitivity analysis can be applied to the signed-rank test easily. The disadvantage on the other hand is that such rank-based tests are geared more to testing for no effect rather than to estimating the effect itself, which results in a roundabout way of getting the point estimate and CI (as shown in the appendix). Since these effect estimates can only be derived under the assumption of homogeneous treatment effects, in contrast with those in Sections 2.1 and 2.2, they are of less interest to the current application. However, the significance level (i.e., the p-value) of the signed-rank test remains valid against an alternative of either homogeneous or heterogeneous treatment effects (and so does the Rosenbaum's sensitivity analysis that follows).

Applying the Wilcoxon (1945) signed-rank test to the current context, rank $|y_{m1} - y_{m2}|$, $m = 1, ..., M$, and denote the resulting ranks as $r_1, ..., r_M$, where a larger rank r_m corresponds to a larger absolute difference in response. The signed-rank statistic is then the sum of the ranks of the pairs where the treated subject has the higher response:

$$R \equiv \sum_{m=1}^{M} r_m 1[s_m = 1].$$

The p-value of the R-statistic can be obtained by the pseudo-sample simulation procedure or the normal approximation method as discussed in Section 2.2. In particular, we show in the appendix that when M is large, the normally approximated p-value for R under the null hypothesis of exchangeability is

$$P(R' \geq R) \simeq P\left\{N(0,1) \geq \frac{R - E(R')}{V(R')^{1/2}}\right\}, \tag{3}$$

where R' is the permuted version of R, $E(R') = M(M+1)/4$, and $V(R') = M(M+1)(2M+1)/24$.

2.4 Sensitivity Analysis with Signed-Rank Test

As noted in Section 2.1, the key identifying assumption for the matching estimator is the 'selection on observables' condition. The same condition is also required for parametric regression approaches. This condition may fail if there are omitted third variables or unobservables that affect both the treatment d (the decision to join the GATT/WTO) and the response y (the trade flows). In a parametric framework, one may deal with this problem of 'selection on unobservables' using techniques such as Heckman's (1979). In the current nonparametric framework, the Rosenbaum (2002) sensitivity analysis provides a convenient way to account for selection on unobservables.

The analysis is structured as follows. Suppose that the treatment d is affected by an unobserved confounder ε. Then, two subjects in a matched pair with the same x but possibly different ε may have different probabilities of taking the treatment. Let the odds ratio of taking the treatment across all matched pairs be bounded between $1/\Gamma$ and Γ for some constant $\Gamma \geq 1$. For instance, if the first subject's probability of taking the treatment is 0.6 and the second subject's 0.5, the odds ratio is $(0.6/0.4)/(0.5/0.5) = 1.5$.

Rosenbaum (2002) shows that given the bounds on the odds ratio, one can derive the corresponding bounds on the significance level of many rank-sum statistics under the null hypothesis of no effect. This places bounds on the significance level that would have been appropriate had ε been observed. The sensitivity analysis for a significance level starts with the scenario of no hidden bias ($\Gamma = 1$). The sensitivity parameter Γ is then increased from 1 to see how the initial conclusion is affected. If it takes a large value of Γ (i.e., a large deviation from 1 in the odds ratio) to eliminate an original finding of a significant effect or to overturn an original finding of no effect, the initial conclusion is deemed robust to unobserved confounders; otherwise, the initial finding is sensitive.

We show in the appendix how to apply the Rosenbaum (2002) sensitivity analysis to the signed-rank statistic and derive the bounds on the significance level (the p-value) of the observed statistic R under the null of no effect. In particular, for a given degree $\Gamma \geq 1$ of departure from the state of no hidden bias, define $p^+ \equiv \frac{\Gamma}{1+\Gamma} \geq 0.5$ and $p^- \equiv \frac{1}{1+\Gamma} \leq 0.5$. The p-value of the observed statistic R is bounded as follows:

$$P(R^+ \geq R) \geq P(R' \geq R) \geq P(R^- \geq R), \tag{4}$$

where $R^+ \equiv \sum_{m=1}^{M} r_m u_m$ with $P(u_m = 1) = p^+$ and $P(u_m = 0) = 1 - p^+$, and likewise for R^-. Note that the means and variances of R^+ and R^- include $E(R')$ and $V(R')$ as a special case when $p^+ = p^- = 1/2$ under no hidden bias.

Specifically, suppose that the H_0-rejection interval is in the upper tail, and the p-value assuming no hidden bias is $P(R' \geq R) = 0.001$, leading to the rejection of H_0 at level $\alpha > 0.001$. By allowing an unobserved confounder to cause the odds ratio to deviate from 1 and up to $(1/\Gamma, \Gamma)$, the correct tail probability is unknown but is bounded above by $P(R^+ \geq R) \simeq P\{N(0,1) \geq \frac{R-E(R^+)}{SD(R^+)}\}$. The upper bound can be obtained for different values of Γ to find the critical value Γ^* at which the upper bound crosses the critical level α.

The relevant distribution (R^+ or R^-) to use for the sensitivity analysis corresponds to the direction of hidden bias that would undermine an initial finding of a significant treatment effect or reverse an initial finding of no effect. Loosely speaking, for example, if the finding is a significantly positive effect, we only need to worry about 'positive' selection, where a subject with a higher potential treatment effect is also more likely to be treated; thus, the relevant distribution is R^+ that embodies selection bias in this direction. On the other hand, if the finding is a significantly negative effect, then 'negative' selection, where a subject with a lower potential treatment effect is also more likely to be treated, can reverse or weaken the original finding; in this case, the sensitivity analysis with R^- is applicable.

As reviewed in the appendix, there exist alternative approaches of sensitivity analysis, but they are typically parametric in nature or not applicable to cases with continuous response variables. In comparison, the Rosenbaum (2002) approach imposes relatively mild assumptions (that the odds ratio of subjects matched on x be bounded between $1/\Gamma$ and Γ) and is straightforward to apply. While most other approaches specify how the unobserved confounder affects both the treatment and response, the Rosenbaum (2002) approach focuses only on how the unobservable may affect the treatment. Thus, the Rosenbaum (2002) approach is likely to be more robust to parametric misspecifications (and at the same time, conservative). On the other hand, by leaving the relationship between the unobserved confounder and the response unspecified, this approach cannot in general construct bias-adjusted effect estimates as in parametric approaches (of the sensitivity analysis nature or of the Heckman type). Instead, this approach evaluates how robust the effect estimate obtained under the assumption of no hidden bias is to the unobserved selection problem. This sensitivity analysis ultimately relies on the researcher's judgement of whether Γ^* at which the initial significance finding reverses is considered large enough. In general, the more important covariates are included in x and the less likely for the odds ratio to be affected by unobserved confounders, the smaller a value for Γ^* can be tolerated. Roughly speaking, we will adopt a threshold of 1.5, which is often adopted by studies using similar sensitivity analysis.[3]

3. DATA DESCRIPTION

We base our analysis on the Rose (2004) data set,[4] although we will also use the Tomz et al. (2007) data set in Section 5.2 as one of the robustness checks. Readers are referred to the source for a detailed account of the data. The data set includes 234,597 observations on trade flows among 178 IMF trading entities between 1948 and 1999 (with some "gaps" and missing observations). There are 12,150 distinct dyads and, on average, about 19 observations for each dyad. The list of variables and their definitions are given in Table 1.

The set of covariates we use for matching are exactly the same as the set of regressors used by Rose (2004) and most other studies in the literature that follow parametric approaches. Thus, we can attribute differences in our findings mainly to the different methodologies taken. In parallel with the previous studies, we will study the effect of GATT/WTO membership, as well as the Generalized System of Preferences (GSP), on a dyad's bilateral trade volume. In particular, two kinds of membership effects are considered: when both countries in a dyad are GATT/WTO members relative to when both are not (both-in effect), and when only one country in a dyad is a GATT/WTO member relative to when both are not (one-in effect). The GSP (which are trade preferences extended from the rich to the poor countries) was found by Rose (2004) to have strong trade effects. We include it in our study to demonstrate that our nonparametric approach can deliver similar effect estimates as the parametric approach in the case of GSP; this provides an anchor to evaluate the drastically different results for both-in and one-in effects.

Table 2 gives the summary statistics of the covariates across three groups of observations by the joint membership status of a dyad (both in, one in, or none in). The control (none-in) dyads on average tend to be closer in distance and smaller in economic sizes, are poorer, and appear in earlier years. Alternatively, based on simple logistic regressions, Table 3 shows that most of the observable covariates affect the selection into membership, and their selection effects (in terms of odds) are statistically significant (different from one). For example, dyads that are farther apart from each other, or larger in economic sizes, are more likely to be GATT/WTO members.

A typical concern about using the matching methods is the extent of overlapping support of the distribution of observable covariates between the treatment and control groups. Figure 1 provides

[3]See Aakvik (2001), Hujer et al. (2004), Caliendo et al. (2005), Hujer and Thomsen (2006), and Lee and Lee (2009), for example.

[4]Available at faculty.haas.berkeley.edu/arose/GATTdataStata.zip

one such visual check often used in the matching literature, where, based on the same logistic regression as above, the propensity score of an observation taking the treatment is estimated and the score's frequencies are tabulated across the treatment and control groups. The histograms in Figure 1 suggest that the supports of the propensity score overlap fairly well between the both-in treated and control groups, or between the one-in treated and control groups.

4. BENCHMARK RESULTS

4.1 Both-In Effects

Table 4 reports the estimation results as we apply the nonparametric methodologies described in Section 2 to the data set of Rose (2004). The both-in effects are significantly positive regardless of the caliper choice (which sets the best 100%, 80%, 60%, or 40% of matched pairs to include in the estimation). The estimates suggest that membership in the GATT/WTO by both countries on average raises bilateral trade volume by 74% ($= e^{0.553} - 1$) to 277% ($= e^{1.328} - 1$) for dyads that both chose to be in the GATT/WTO. In contrast, bilateral trade volumes would have increased by 20% ($= e^{0.185} - 1$) to 40% ($= e^{0.337} - 1$) if the nonmember dyads had both joined the GATT/WTO. The both-in effect on all is positive and significant, reflecting in large part the effect on the treated.

The significant difference between the both-in effect on the treated and untreated suggests the presence of heterogeneous treatment effects. To see this, note that if the treatment effect is homogeneous regardless of x, then we do not need to worry about the separate effect on the treated and untreated, as they should be the same. However, if the effects are heterogeneous and vary with x, and if the selection into the treatment also depends on x (as the previous section showed) such that x is on average different between the treatment and control groups, then the effect on the treated and untreated will be different.

The findings are very similar when the estimation is based on the signed-rank test (the R-statistics) instead of the original permutation test (the D-statistics). This remains the case throughout our analysis. Thus, we will focus on the effect estimates based on the D-statistics that theoretically allows for heterogeneous effects. Nonetheless, the p-value of the signed-rank test will be focal, because it is the basis for the Rosenbaum (2002) sensitivity analysis.

Results of the sensitivity analysis indicate that the positive both-in effect on the treated is robust to selection bias to the extent that a treated subject is not 2.081 times (and beyond) more likely than a comparable untreated subject to take the treatment (by the 80% caliper and the two-sided test). The robustness ranges from 1.467 to 2.434 as the test or the caliper choice varies. By the threshold of 1.5, the above finding is reasonably robust. In comparison, the both-in effect on the untreated is less robust to potential hidden bias. Overall, we see strong evidence for a positive *realized* both-in effect on member dyads (and less so for a positive *potential* effect on nonmember dyads).

On theoretical grounds, several economic models predict a positive both-in effect on trade. Among others, the terms-of-trade argument (Johnson, 1953–1954; Bagwell and Staiger, 1999, 2001) suggests that multilateral trade agreements help coordinate countries' trade policies and remove their terms-of-trade incentives to raise trade barriers. The terms-of-trade incentive is shown by Broda et al. (2008) to be an important factor indeed in non-WTO countries' trade policy. The political-commitment argument (Staiger and Tabellini, 1987, 1989, 1999), on the other hand, suggests that multilateral trade agreements help national governments commit themselves to liberalized trade policies, bringing about efficient production and trade structures.

In spite of the above theories, there are several empirical difficulties in using membership to measure the GATT/WTO effect, as noted by many in the literature, cf. Rose (2010). First, tariff reductions and policy liberalizations do not necessarily coincide with the date of accession. Second, some GATT/WTO members may extend their most-favored-nation (MFN) treatment to nonmember trading partners. Third, some countries (particularly developing countries) did not liberalize their trade policies in spite of their membership in the GATT (although this is less the case under the WTO). Fourth, some sectors (e.g., oils and minerals) face little protectionism with or without the GATT/WTO, while some (e.g., agriculture) are highly protected with or without the GATT/WTO. The first two considerations imply that membership is a noisy measure (as a result, the estimates will be downward biased), while the last two imply that GATT/WTO effect is heterogeneous with no effect in some cases. The fact that we obtained positive significant effects implies that on average across many trading relationships, the theoretical both-in effect is strong enough to dominate the above factors and to leave an empirically measurable impact.

4.2 One-In and GSP Effects

Unlike the both-in effect where one may expect a positive effect, or a zero effect at worst, *a priori*, the one-in effect can take either sign. On one hand, import diversion by the new member from its nonmember trading partner to other member trading partners may lower the dyad's bilateral trade volumes. On the other hand, in many cases, when a country joins the GATT/WTO, its tariff reductions (and other policy liberalizations) offered to members on a MFN basis are also extended to nonmember trading partners. In this case, imports increase from all sources, including nonmember trading partners. Furthermore, when a country gains access to the markets of existing GATT/WTO members with the newly acquired membership, it may increase imports of inputs necessary for the production of exports to these destinations. Some of these additional imports may fall on third nonmember countries. For example, with the accession into WTO, China may increase imports of oil from Iran in its expansion of production and export activities.

The results in Table 4 suggest that the one-in effect on the treated is overall positive and significant: the estimates range from 39% ($= e^{0.326} - 1$) to 115% ($= e^{0.767} - 1$). Thus, it appears that the trade-creating effects dominate the potential trade-diverting effects, for dyads where one country has unilaterally joined the GATT/WTO. Similar to the both-in effect, the one-in effect on the untreated is smaller and less robust to potential hidden bias. Although there are exceptions in our following analysis, overall, the evidence for a positive GATT/WTO effect on the untreated is not strong (we may say that countries have selected well in the sense that they only joined the GATT/WTO if the perceived benefits were large). Thus, we will report only the effect on the treated in what follows. An extended set of estimates are available in an unabridged version (Chang and Lee, 2010) of this paper.

The GSP scheme is also found to promote bilateral trade, by a factor of 94% ($= e^{0.665} - 1$) to 134% ($= e^{0.851} - 1$) [the upper bound estimate is very close to Rose's (2004) benchmark estimate 136% ($= e^{0.86} - 1$)].[5] The GSP effect estimates are smaller than the both-in effects, but larger than the one-in effects in general. This ranking seems to make sense in theory. As the GSP is a system of unilateral trade preferences extended only from a high-income country to its poor trading partners, its likely effect on bilateral trade volumes is *a priori* smaller than if both the rich and the poor countries in a dyad lower their import restrictions against each other, which happens presumably if both join the GATT/WTO. On the other hand, any trade-promoting effect of the one-in membership is, as argued above, indirect and conditional on the spillover of the MFN treatment and on the dyad's initial trade pattern, while the effect of GSP is directly derived from a straightforward reduction of dyad-specific trade resistance.

It may be helpful to point out that the positive and stronger trade effect of both-in is shared by a larger number of bilateral trading relationships (114, 750) than that of GSP (54, 285). Thus, either on the average or in the aggregate, our estimation results suggest that the realized trade-creating effect of GATT/WTO membership is larger than GSP.

5. ROBUSTNESS CHECKS

In this section, we conduct an extensive set of robustness checks by considering various restricted matching criteria, alternative GATT/WTO indicators, the non-randomness of zero trade flows, the inclusion of multilateral resistance terms, and different matching methodologies. Overall, the benchmark finding of a significant GATT/WTO effect on trade is strengthened, not weakened, while the GSP effect is qualified.

5.1 Restricted Matching

Although we did the Rosenbaum (2002) analysis to assess the sensitivity of the benchmark results to whatever selection bias may remain after controlling for x, the analysis itself does not remove the bias. In the literature, three potential sources of bias seem to be of major concern. They are systematic unobservable heterogeneity across dyads, years, and development stages that may influence bilateral trade volumes as well as selection into GATT/WTO. In view of this, we restrict the potential match for a subject to observations that have the opposite treatment status (as in the benchmark case) and are furthermore from the same dyad, the same year, or the same relative development stage, alternately. By doing this, we control for the likely dyad, year, or development-stage specific effect.[6]

[5] We did not report the GSP effect on the untreated, as the GSP does not apply to all kinds of trading relationships. For example, it is not relevant to propose a GSP between two poor countries.

[6] In Chang and Lee (2010), we also conduct restricted matching within the same time period, with the periods defined according to the GATT/WTO trade negotiation rounds. The estimates are almost the same as in unrestricted

Table 5 summarizes the restricted matching results (we repeat in the first sub-column the relevant information from the benchmark case for ease of comparison). The estimates suggest that the positive both-in effect continues to be economically and statistically significant, and larger than either the one-in or GSP effect.[7] In contrast with the 'within-year' estimates that measure cross-sectional (or 'between') variations, the 'within-dyad' estimates measure time-series (or 'within') variations. Both 'within' and 'between' variations indicate that there are significant gains in trade volumes by joining the GATT/WTO.

Note that the 'within-year' estimates are almost identical to the benchmark results. This indicates that in unrestricted matching, the matched subjects are often from the same year; thus, the benchmark estimates pick up mostly cross-sectional variations. This is understandable, as the set of covariates include year dummies, which encourages matching observations from the same year. A further look into the data (not reported in the table) at every five-year interval $(1950, 1955, \ldots, 1995)$ shows that the positive both-in or one-in effect is not lumpy in a few particular years but is felt throughout the years, except in 1975 and 1995 when there is a dip in the membership effects.

The 'within-devel.' analysis reports results when matching is restricted to the same development stage combination, where the combinations are: low-income/low-income, low-income/middle-income, low-income/high-income, middle-income/middle-income, middle-income/high-income, and high-income/high-income dyads. Are the positive membership effects shared evenly among countries of different development stages, or are they concentrated on particular subsets of countries? A look into the data (not reported in the table) shows that the positive effects are indeed concentrated on dyads of middle-income/middle-income, middle-income/high-income, and high-income/high-income countries. The low-income countries do not benefit much from a membership in the GATT/WTO. Similar lumpy patterns were found in Subramanian and Wei (2007), although we still find a positive average effect while they found no positive average effect.

This asymmetry may reflect the two empirical concerns mentioned above: that the low-income countries do not significantly liberalize their import sectors despite their membership in the GATT/WTO and that major export sectors (e.g., agriculture) of low-income countries still face steep protectionism from the rich world with or without the GATT/WTO. This kind of heterogeneity in GATT/WTO membership effects is implied by existing theories of trade agreements; see, for example, Bagwell and Staiger (2010, pp. 245–247) for a review. Basically, the two GATT/WTO principles of MFN and reciprocity actually facilitate this outcome, whereby if countries do not actively participate in trade negotiations/tariff reductions, other active players can engineer tariff bargains among themselves that minimize free-riding by third countries. Thus, by not offering

matching, which is not surprising, given our finding below that matched subjects in unrestricted matching often come from the same year; thus, the criterion of matching within period does not impose extra restriction in most cases.

[7] The number of matched pairs obtained when matching is restricted within the same dyad shrinks substantially, as some dyads may not have both treated and untreated observations during the sampling years. For example, the 'US-Japan' dyad has 'one-in' (years 1950–1954) and 'both-in' (years 1955–1999) observations but does not have 'none-in' observations. In cases like this, dyads without qualified control/treated subjects are dropped from the estimation.

domestic market access, the low-income countries may also face difficulty expanding their export volumes.

5.2 Participation versus Formal Membership

Tomz et al. (2007) stress the importance of *de facto* participation in the multilateral system by nonmembers such as colonies, newly independent colonies, and provisional members. They share to a large extent the same set of rights and obligations under the agreement as formal members. Tomz et al. (2007) classify these territories as nonmember participants and define participation to include both formal membership and nonmember participation. Based on the same estimation framework of Rose (2004), they find significant participation effects on trade.

Table 6 reports the nonparametric estimates given the data set of Tomz et al. (2007) and the alternative GATT/WTO indicator. We see that participation effects are overall stronger than membership effects reported earlier; they are also more robust to hidden selection bias. This finding of a larger participation than membership effect is consistent with the contrasting results reported by Tomz et al. (2007) and Rose (2004).[8]

5.3 Kernel-Weighting Matching versus Pair Matching

In contrast with pair matching, which uses only the nearest match, kernel-weighting matching uses multiple potential matches by attaching greater weights to nearer matches. The weighting scheme depends on the chosen kernel and bandwidth. In this exercise, we use the normal kernel and define weights for the potential matches $i't'$ of a subject it as $w_{it,i't'} \equiv \phi(\frac{x_{1,it}-x_{1,i't'}}{SD(x_1)h}) \dots \phi(\frac{x_{P,it}-x_{P,i't'}}{SD(x_P)h})$ where $\phi(\cdot)$ denotes the standard normal density function, P the dimension of the covariate vector x, $SD(x_p)$ the standard deviation of a covariate x_p in the pooled sample, and h the chosen bandwidth.[9] The kernel-weighting matching estimator is then defined as $\frac{1}{M}\sum_{it}(y_{it} - \sum_{i't'} \tilde{w}_{it,i't'} y_{i't'})$, where $\tilde{w}_{it,i't'} \equiv w_{it,i't'} / \sum_{i't'} w_{it,i't'}$ is the normalized weight. Table 7 summarizes the results. The effect estimates are very similar to those obtained by pair matching across types of treatments, calipers, and the matching criteria.[10]

5.4 Non-random Incidence of Positive Trade Flows

By using the data set of Rose (2004) or Tomz et al. (2007), we have based our analysis on observations with positive trade flows. Recent studies by Helpman et al. (2008) and Felbermayr

[8]As shown in the table, the GSP effect estimates are not exactly the same as those based on the Rose (2004) data set, for two reasons: first, when the GSP effect is estimated, the participation status of a dyad replaces membership status as part of the covariates. Second, Tomz et al. (2007) also corrected some coding errors in Rose's data set, in particular, the income status and geography indicator of some territories (Tomz et al., 2007, Foonote 32). The second reason also explains the difference in the number of matched pairs obtained for GSP under 'within-devel.' with the alternative data set.

[9]For matching within dyad where the number N_{it} of potential comparison subjects for a subject it is small, we use a larger bandwidth $h = 0.5N_{it}^{-1/(P+4)}$; otherwise, we use a smaller bandwidth $h = 0.25N_{it}^{-1/(P+4)}$ (the computation hits numerical bounds for smaller bandwidths than this).

[10]We set calipers in the same fashion as in pair matching, such that subject it that does not have a good match in terms of the scale-normalized distance is discarded. We also experiment with larger bandwidths. As the chosen bandwidth is enlarged, the point effect estimates tend to increase. Thus, we may consider the pair matching estimates as overall conservative estimates.

and Kohler (2007) stress the importance of incorporating observations with zero trade flows in estimating the gravity equation. In particular, both studies find that GATT/WTO membership has a positive effect on the formation of bilateral trading relationships. This suggests that using only observations with positive trade flows will induce a downward bias in the effect estimate of GATT/WTO membership (and other trade barriers as well), since a pair of countries that are not GATT/WTO members but still observed trading with each other are likely to have lower unobserved trade resistance. Both studies find that consideration of this selection bias alone indeed strengthens the gravity equation estimates, albeit not considerably.[11]

Given that we found a strong and positive membership effect based on positive trade flows, the above selection argument suggests that incorporating observations with zero trade flows in our analysis will only strengthen the initial finding of a positive effect. Thus, we do not expect our general conclusions to change with the inclusion of zero trade. Both studies by Helpman et al. (2008) and Felbermayr and Kohler (2007) are based on parametric estimations of the trade flow equation, although the former considers parametric as well as nonparametric estimation of the selection equation. To estimate the membership effect and also to address the selection into trading in a fully nonparametric framework, one can potentially apply the newly proposed methodology of Lee (2010). We leave this considerably more extensive work for future research, and attempt a less ambitious approach here to isolating the GATT/WTO membership effect on trade volumes from its effect on 'trade start' without resorting to a new data set and a full-blown new estimation framework.

Still based on the Rose (2004) data set, we use only observations where the two countries in a dyad start trading with each other before ever joining the GATT/WTO. In other words, these dyads have reported bilateral trade flows before either one of them ever joins the GATT/WTO. Using this sub-sample of dyads that trade with or without the GATT/WTO membership, the membership effect on prompting new trading relationships is not present; thus, the effect estimates consist only of the membership effect on trade volumes. Table 8 presents the effect estimates for this sub-sample following the same matching procedure as in the benchmark and restricted cases. We see that this refined analysis reports overall stronger membership effects, and thus in a way the results are consistent with the above selection argument.

5.5 Multilateral Resistance

Relative trade resistance rather than absolute trade resistance is argued by some gravity theories to be more appropriate in explaining bilateral trade flows, cf. Anderson and van Wincoop (2003), and thus multilateral resistance (MR) terms may have to be controlled for. As their paper suggested, there are two ways to control for the terms. One is to solve the endogenous MR terms given the

[11] Helpman et al. (2008) also distinguish the direct partial effect of trade resistance on trade flows from its indirect effect on trade flows through changes in the number of exporters. In this paper, we have not made this distinction. In our view, the larger trade flows due to an increase in the number of exporters should also be considered as part of the benefit of GATT/WTO membership. Thus, the matching estimates presented correspond to the total effect of GATT/WTO membership, including both the direct and indirect effects.

parameter values and then to estimate the parametric gravity equation incorporating dyads' MR terms by nonlinear least squares. Both the solution to the endogenous MR terms and the parametric gravity equation rely on certain functional form assumptions and thus are subject to specification errors as noted by the authors themselves, which are exactly what we try to avoid in the current paper. An alternative suggested by the same authors is to replace the MR terms with country dummies. In a way, we have controlled for dyad-specific and hence country-specific effects when we conduct the matching within the same dyad; the strong effects of GATT/WTO remained. On the other hand, we do not have a good way in the matching framework to control for time-varying country-specific effects as emphasized by some parametric studies, cf. Subramanian and Wei (2007).

Recent studies by Baier and Bergstrand (2009a,b) present some potential methods to approximate the endogenous MR terms by observable exogenous trade resistance covariates and thus the possibilities to address time-varying MR terms in the matching framework. Specifically, in one version of their proposed approximations, the two country-specific MR terms for a dyad are decomposed into a list of MR terms associated with each trade resistance covariate. For example, the MR term for a trade resistance covariate x^r_{kmt} between countries k and m in year t would be $MRx^r_{kmt} = (1/N)\sum_{m'=1}^{N} x^r_{km't} + (1/N)\sum_{k'=1}^{N} x^r_{k'mt} - (1/N^2)\sum_{k'=1}^{N}\sum_{m'=1}^{N} x^r_{k'm't}$, reflecting the respective average trade resistance of the two countries to all their trading partners, adjusted by a typical country's average resistance to all its trading partners. One can add this list of MR terms to the list of covariates already used in the matching.[12] Specifically, to estimate the both-in treatment effect, we follow the same matching procedure as in the benchmark case but with the modified list of matching covariates that include the same economic size covariates (*lrgdp*, *lrgdppc*, *lareap*), the trade resistance covariates (*ldist*, *comlang*, ..., *regional*, *gsp*) and their corresponding MR terms, year dummies, and the MR term of the treatment dummy.[13] Similar adjustments are made to the list of matching covariates for one-in and GSP effect estimation.

The results are summarized in Table 9. When the multilateral resistance terms are controlled for, we see that the strong both-in effects on the treated remain. In contrast, the one-in treatment effects now become weaker overall with statistically significant but small trade promoting effects. The 'within-year' matching results are almost identical to the unrestricted case, reflecting again the fact that in the unrestricted case, most matched observations are across sections from the same year. The 'within-dyad' estimates of the both-in and one-in effects are comparable to those in Table 5 without the MR terms controlled for. This suggests that the MR terms do not vary much across years for a given dyad, and hence the extra control does not affect the matching significantly. When matching is restricted within the same relative development stage, the mean both-in effect again

[12]Alternatively, one can construct the relative trade resistance covariate $BVx^r_{kmt} \equiv x^r_{kmt} - MRx^r_{kmt}$ and use it in place of the absolute trade resistance covariate x^r_{kmt} in the matching, as done in Baier and Bergstrand (2009b). We take the former approach, as it imposes less structure.

[13]Note that we have included the MR term of *bothin* in the list of matching covariates in estimating the both-in treatment effect; thus, the estimated both-in effect corresponds to its partial equilibrium effect and not its potential general equilibrium effect (the estimation of which goes against the typical assumption of matching estimation). In the context of free trade agreements (FTAs) that Baier and Bergstrand (2009b) studied, they argued that the effect of the MR term of their treatment dummy, FTA, was conceptually negligible.

masks a large variation across dyads of different development stages (not reported in the table), with large benefits tending to concentrate on higher income dyads but costs on lower income dyads.[14]

5.6 Difference-in-Difference Matching Estimator

In this section, we explore an alternative treatment effect concept, difference-in-difference (DD), which is based on weaker identification assumptions. This method compares the difference over time in trade volumes of a treated dyad to that of a comparable untreated dyad. Consider a time period $[t-b, t+a]$ around the treatment timing t with $a, b > 0$. Using our notations, the DD treatment effect estimand is:

$$\begin{aligned} DD &= E(y_{t+a} - y_{t-b}|d=1, x) - E(y_{t+a} - y_{t-b}|d=0, x) \\ &= E(y^1_{t+a} - y^0_{t-b}|d=1, x) - E(y^0_{t+a} - y^0_{t-b}|d=0, x) \\ &= E(y^1_{t+a} - y^0_{t+a}|d=1, x) \end{aligned} \tag{5}$$

if the *same time-effect* condition $E(y^0_{t+a} - y^0_{t-b}|d=1, x) = E(y^0_{t+a} - y^0_{t-b}|d=0, x)$ holds. That is, DD identifies *the treatment effect on the treated at time* $t+a$ if the potential untreated response changes by the same magnitude on average over the time period $[t-b, t+a]$ for comparable treated and untreated dyads. This identifying assumption is weaker than $E(y^0|d=1, x) = E(y^0|d=0, x)$ required for the effect on the treated (cf. Section 2.1) and thus is more robust to hidden bias due to selection on unobservables. For example, the same time-effect condition allows potential systematic unobserved dyadic heterogeneities across the treatment and control groups or systematic time trends in trade volumes unrelated to the treatment, as long as the time trends are on average the same for comparable dyads. See Heckman et al. (1997) for DD estimation based on matching, and Imbens and Wooldridge (2009) and the references therein for other DD approaches.

To estimate DD, we carry out matching in a fashion similar to Section 2.1. In particular, start with a both-in treated dyad. If the dyad was first treated in year t, the pool of potential matches for this dyad are dyads that were not in the GATT/WTO throughout the period $[t-b, t+a]$. The best match is identified based on the baseline response and the covariates in the pre-treatment year (y_{t-b}, x_{t-b}).[15] Given the match, the difference over time in trade flows $(y^0_{t+a} - y^0_{t-b})$ of the control dyad is subtracted from the difference over time $(y^1_{t+a} - y^0_{t-b})$ of the treated dyad. Given M matched pairs, DD is estimated by the sample average of the pair-wise differences in differences. The one-in and the GSP treatment analysis are carried out in a similar fashion. Note that we have included the baseline response y_{t-b} in the list of matching covariates. This is to control for potential unobservables that may systematically affect trade flows but are not captured by the observables x_{t-b}, and thus to reduce the scope of selection on unobservables.

Some remarks are in order. First, selecting the lead and lag years (a, b) is difficult. One guideline

[14]The GSP treatment effects are stronger with the MR terms controlled for as in the case of both-in effects.

[15]The same scale-normalized distance measure is used, with the sample variance of (y_{t-b}, x_{t-b}) calculated based on all observations in year $t-b$.

is whether the same time effect condition will hold given the choice of (a, b). As noted earlier, policy changes do not necessarily coincide with the official year of GATT/WTO accession. Some countries may undertake structural changes required for the accession beforehand or economic agents may act in anticipation of the upcoming accession. Thus, trade flows may well have changed before the official accession of the treated dyad, and to satisfy the same time effect condition, a large b may be required. On the other hand, it is quite often true that acceding countries take several years to phase in the agreed-upon trade policy changes, and thus one may expect the treatment effect to manifest itself only years later. A large a may address this concern. However, choosing too large a window (a, b) may pose two problems: first, the sample size will be significantly reduced as not all dyads have observations in long extended periods; second, with a long window, other factors not controlled for (by the same time effect condition and the matching covariates) may affect the trade flows and contaminate the result. We experiment with several symmetric windows: $a = b = \{1, 2, \ldots, 6\}$. Another remark worth making is that a dyad typically went from a none-in period to a one-in period and then to a both-in period, if it was ever both-in treated. It is relatively rare for the countries in a dyad to simultaneously join the GATT/WTO and to go directly from none-in to both-in. To maintain reasonable sample sizes, we allow both scenarios of pre-treatment status (none-in or one-in) in estimating the both-in treatment effect. Thus, the both-in effect estimate is a mixture of the two effects when the dyad goes from one-in to both-in and when the dyad goes from none-in to both-in, relative to if it stays none-in throughout the interval. The one-in and the GSP effect analysis are spared such complications.

The findings are summarized in Figure 2. The results are similar across different caliper choices. In general, the GATT/WTO trade effects are negligible in early phases of the membership, but become statistically and economically significant five or six years after the treatment. At year six, an average dyad's bilateral trade flows increase roughly by 65% ($= e^{0.5} - 1$). Similar patterns apply to the both-in or one-in treatment. In contrast, the GSP effect is small if not negligible and manifests itself relatively quickly following the treatment. The effect remains relatively stable throughout the years, and is statistically insignificant in most cases.

These findings seem to agree with the casual observations and our discussions above regarding the gradual phase-in of policy changes after an official GATT/WTO accession. It may also be reconcilable with the larger benchmark and restricted matching estimates shown in Tables 4 and 5. In these earlier exercises, we did not control for the vintage of the treated observations; thus, the treatment effect estimate effectively summarizes the effects across all vintages following the treatment for as far as several decades. If the effect is larger, the more aged the treatment is, a larger effect estimate observed in the previous exercises is understandable.

5.6.1 Placebo Exercise. In this section, we conduct "placebo" exercises to verify that the time trends of trade flows of matched dyads are comparable in advance of membership. A finding against differences in pre-trends would help alleviate concerns that the DD estimates may be picking up systematic differences in time trends between the treatment and control groups due to unobservables not controlled for in our matching exercise. To do so, we apply the DD estimation procedure to a bogus treatment year $t' = t - d$ that predates the actual year of treatment t (here identified as the first year when either country in a treated dyad joins the GATT/WTO). As there is no treatment at the bogus treatment year, the DD estimate, instead of estimating the treatment effect, captures the difference in the time trends between comparable treated and untreated dyads in advance of GATT/WTO membership.

As discussed above, countries may undertake policy reforms in advance of membership, and their trade patterns may well have changed years before the official year of accession. Thus, the period of comparison of the pre-trends has to be set reasonably far into the past, such that it does not overlap with the likely period of transition to the accession. For this, we experiment with $d = \{7, \ldots, 12\}$ and symmetric DD windows $a = b = \{1, \ldots, 6\}$, with $d - a \geq 6$. That is, the period of comparison of the pre-trends will be at least six years before the actual year of treatment. For example, if the bogus treatment year is set 10 years before the actual treatment year, the forward/backward window for DD estimation can range from one to four years.

The results are summarized in Table 10. As can be seen from the table, of the 21 possible periods of comparison (and of the four caliper choices for each period), all estimates are not significantly different from zero, except three estimates that are significantly negative (which does not go against a finding of positive treatment effects). Thus, on the whole, there is no evidence of systematic differences in the time trends in advance of membership between comparable treated and untreated dyads.

6. POTENTIAL PROBLEMS WITH THE PARAMETRIC GRAVITY ESTIMATES

The discrepancy between the current nonparametric matching estimates and the conventional parametric gravity estimates suggests that the empirical gravity models used in the parametric studies may be misspecified. In particular, guided by the pattern of nonparametric effect estimates observed above, we suspect that heterogeneous treatment effects can be important. While our matching estimator allows for heterogeneous effects that vary with the observed covariates, the specifications used in Rose (2004) basically assume homogeneous GATT/WTO effects. Subramanian and Wei (2007) allow for heterogeneous effects in the parametric framework but only across certain subsets of samples. In this section, we explore generalizing the parametric gravity model to allow for more arbitrary forms of heterogeneous effects and verify whether the discrepancy in findings between the nonparametric and parametric approaches might be reduced. To work toward this, we introduce first-order interaction terms of the GATT/WTO indicators with the other covariates.[16]

[16] We also explore adding quadratic terms of continuous/categorical covariates and interactions of these covariates with all other binary covariates (other than the treatment variables themselves) to the Rose (2004) default specification. Many of these terms are significant, but the OLS estimates of the membership effects are not affected significantly.

The results are summarized in Table 11. As shown, when only the *bothin* GATT/WTO indicator is allowed to interact with the other covariates, the general finding does not change, although many of the interaction terms are significant. As both the *bothin* and *onein* GATT/WTO indicators are allowed to interact with the other covariates, the mean effects of both membership statuses become significantly positive. Many of the interaction terms are statistically significant, and the default model is rejected in favor of the alternative model. While the estimates for the main gravity covariates (such as distance and GDP) remain stable across specifications, estimates for the other covariates are not, suggesting that the modeling of these augmenting covariates (typically used to control for the degree of trade resistance) is problematic. Basically, the parametric effect estimates of these augmenting trade resistance covariates are very sensitive to the model specifications. This may help explain some of the disagreements in the gravity literature regarding the currency union effect (Persson, 2001; Rose, 2001) or the free trade agreement effect (Frankel, 1997; Baier and Bergstrand, 2007).

Finally, as the *gsp* dummy is also allowed to interact with the other covariates, the mean effect estimates of the both-in and the one-in membership status remain significantly positive. The GSP mean effect estimate is, however, rather similar to its marginal effect estimate in the default specification. This suggests that allowing for heterogeneous GSP effects helps in increasing the explanatory power of the model but the degree of heterogeneity is not strong, compared with the both-in and one-in effects. This also agrees with the findings of the matching framework above: while the GSP effect estimates are relatively stable across the choice of calipers, the both-in and one-in effect estimates vary a lot, and while the GSP effect estimates are relatively similar across the parametric and nonparametric approaches, the membership effect estimates are very different across the two approaches.

Based on the results in the last column of Table 11, it appears that the both-in and one-in membership effects are intensified by the GDP per capita and the physical areas of the dyad, and are also intensified if the countries in a dyad share a common language, were ever in a colonial relationship, or belong to a common currency union. Overall, the explorations above suggest that it is important in practice to recognize the potential heterogeneity in the trade effects of GATT/WTO membership.

In the Rose (2004) default specification, the MR terms are not controlled for. We also explored controlling for the MR terms before proceeding with the same experiment as above of adding interaction terms. In particular, we follow Subramanian and Wei (2007) and use time-varying country dummies as proxies for the MR terms in the Rose (2004) parametric framework.[17] The findings are similar to those above without the MR terms controlled for. The both-in and one-in effect estimates are not statistically significant by controlling for the MR terms alone. By

[17] Instead of using the complete Rose (2004) data set, only observations at every five years between 1950 and 1995 are used. This is to keep the number of time-varying country dummies computationally manageable; see Subramanian and Wei (2007) for the same approach. Five variables—*lrgdp*, *lrgdppc*, *landl*, *island*, *lareap*—are dropped from the list of regressors, as their coefficients cannot be precisely estimated with the presence of time-varying country dummies; their higher-order terms or interaction terms with the other covariates can still be included, however.

incorporating the interaction terms of the membership indicators with the other covariates, the effects turn significantly positive. The set of statistically significant interaction terms are similar: e.g., GDP per capita, a common language, and having been in a colonial relationship tend to strengthen the membership effects.

As the dimension of the covariate vector is high in the current application, there are many potential functional forms for the interaction terms. For example, the GATT/WTO indicators may also interact with the interaction terms of the other covariates (this is where nonparametric methods come in particularly useful; nonparametric methods deliver findings without the need to search for the correct specification). By considering only the first-order interaction terms, we have stopped short of fully explaining away the discrepancy between the effect estimates of the nonparametric and parametric approaches. Nonetheless, our limited search suggests that the assumption of homogeneous treatment effects could be a major source of misspecification. The nonparametric framework we propose in this paper offers a convenient estimation framework to accommodate heterogeneous treatment effects and at the same time circumvents the specification difficulty in a high-dimensional application.

7. CONCLUSION

This paper contributes to the literature on the effects of GATT/WTO membership/participation on actual trade flows. Previous studies of this issue have largely relied on parametric estimation of gravity-based trade models. Concerns about parametric misspecifications, heterogeneous membership effects, and unobserved selection bias are raised by the current paper and addressed by using nonparametric methods. In particular, a pair-matching estimator is used to obtain the point effect estimates, permutation tests to derive the inferences, and a sensitivity analysis based on signed-rank tests to evaluate the robustness of the inferences to unobserved confounders.

Our findings suggest that membership in the GATT/WTO has a significant trade-promoting effect for dyads that have both chosen to be members. The effect is larger than bilateral trade preference arrangements, GSP, and larger than when only one country in a dyad has chosen to be a member. Although the GSP effect appears to be relatively constant across subjects, the both-in and one-in effects display substantial heterogeneities. The finding of a positive both-in effect is quite robust to potential unobserved confounders.

The overall conclusion does not change when we restrict the matching to observations from the same dyad (thus, capturing the within effect), the same year (thus, capturing the between effect), or the same relative development stage. The overall conclusion does not change either when we use participation status instead of formal membership as the treatment indicator, or when we use kernel-weighting matching instead of pair-matching. The results are also robust to using only observations where a dyad's trading relationship exists before either country in the dyad ever joins the GATT/WTO (thus, isolating the membership's effect on trade volumes from its effect on the formation of trading relationships), and robust to controlling for time-varying multilateral resistance terms in the matching framework. A final robustness check using the difference-in-difference matching estimator reveals that the significant and positive GATT/WTO effect on trade takes several years after the official accession before manifesting itself.

The contrast between the results of the current paper and those of Rose (2004) suggests that conventional gravity models may be misspecified. We show that the assumption of homogeneous membership effects may be a major source of misspecification. The nonparametric framework we propose in this paper offers a convenient estimation framework to accommodate heterogeneous treatment effects and at the same time circumvents the specification difficulty in a high-dimensional application.

8. APPENDIX: PERMUTATION TEST FOR MATCHED PAIRS

Recall that $D' \equiv \frac{1}{M}\sum_{m=1}^{M} w_m s_m (y_{m1} - y_{m2})$, where only the permutation variable w_m is random with $P(w_m = 1) = P(w_m = -1) = 0.5$, conditional on the observed data. Hence, $E(D') = 0$ and $V(D') = E(D'^2) = \frac{1}{M^2}\sum_{m=1}^{M} E\{w_m^2 s_m^2 (y_{m1} - y_{m2})^2\} = \frac{1}{M^2}\sum_{m=1}^{M}(y_{m1} - y_{m2})^2$. By applying the central limit theorem to w_m's, the exact p-value of D can be approximated by

$$\begin{aligned}
P(D' \geq D) &= P\left\{\frac{D'}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}} \geq \frac{D}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}}\right\} \\
&\simeq P\left\{N(0,1) \geq \frac{D}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}}\right\}.
\end{aligned}$$

We can obtain the CI for the mean effect by inverting the above test procedure. For instance, suppose that the treatment effect is β_m for pair m. Define the mean effect $\bar{\beta} \equiv \frac{1}{M}\sum_{m=1}^{M}\beta_m$. In this case, the no-effect situation is restored by replacing y_{m1} with $y_{m1} - \beta_m$ when $s_m = 1$ or y_{m2} with $y_{m2} - \beta_m$ when $s_m = -1$:

$$\begin{aligned}
D_{\bar{\beta}} \equiv \frac{1}{M}\sum_{m=1}^{M} s_m(y_{m1} - s_m\beta_m - y_{m2}) &= \frac{1}{M}\sum_{m=1}^{M} s_m(y_{m1} - y_{m2}) - \frac{1}{M}\sum_{m=1}^{M}\beta_m \\
&= \frac{1}{M}\sum_{m=1}^{M} s_m(y_{m1} - y_{m2}) - \bar{\beta},
\end{aligned}$$

and the permutation test can be applied. Define accordingly $D'_{\bar{\beta}} \equiv \frac{1}{M}\sum_{m=1}^{M} w_m[s_m(y_{m1} - y_{m2}) - \bar{\beta}]$ to observe $E(D'_{\bar{\beta}}) = 0$ and $V(D'_{\bar{\beta}}) = \frac{1}{M^2}\sum_{m=1}^{M}[s_m(y_{m1} - y_{m2}) - \bar{\beta}]^2$. Now conduct level-$\alpha$ tests with

$$\frac{D_{\bar{\beta}}}{\{\sum_{m=1}^{M}[s_m(y_{m1} - y_{m2}) - \bar{\beta}]^2/M^2\}^{1/2}}. \tag{6}$$

The collection of $\bar{\beta}$ values that are not rejected using (6) is the $(1-\alpha)100\%$ CI for $\bar{\beta}$. In the above framework, we have generalized the procedure to allow for heterogeneous treatment effects, and as such, the CI constructed is for the mean effect $\bar{\beta}$. Clearly, this framework includes homogeneous treatment effects as a special case when $\beta_m = \beta$ for all m.

9. APPENDIX: SIGNED-RANK TEST FOR MATCHED PAIRS

The permuted version R' for R can be written as $R' \equiv \sum_{m=1}^{M} r_m 1[w_m s_m > 0] = \sum_{m=1}^{M} r_m (1[w_m = 1, s_m = 1] + 1[w_m = -1, s_m = -1])$. Note that r_m's and s_m's are fixed conditional on the data and the only thing random is the permutation variable w_m. Thus, under the H_0 of exchangeability, $E(R') = \sum_{m=1}^{M} r_m/2 = M(M+1)/4$, and $V(R') = \sum_{m=1}^{M} r_m^2/4 = M(M+1)(2M+1)/24$. Hence, when M is large, the normally approximated p-value for R is

$$P\left\{N(0,1) \geq \frac{R \;-\; M(M+1)/4}{\{M(M+1)(2M+1)/24\}^{1/2}}\right\}.$$

Under the assumption of homogeneous treatment effects, the CI for the effect can be obtained by inverting the signed-rank test procedure. Conduct level-α tests with different values of β using

$$\frac{R_\beta \;-\; M(M+1)/4}{\{M(M+1)(2M+1)/24\}^{1/2}}, \quad \text{where} \quad R_\beta \equiv \sum_{m=1}^{M} r_{m\beta} 1[s_m(y_{m1} - s_m\beta - y_{m2}) > 0] \tag{7}$$

and $r_{m\beta}$ is the rank of $|y_{m1} - s_m\beta - y_{m2}|$, $m = 1, ..., M$. The collection of β values that are not rejected is the $(1-\alpha)100\%$ CI for β. To obtain a point estimate of the treatment effect, we can use the Hodges and Lehmann (1963) estimator, which is the solution of β such that

$$R_\beta = \frac{M(M+1)}{4} \; \{= E(R')\}. \tag{8}$$

Note that when treatment effects are heterogeneous, the pair-wise effect β_m (instead of β) should be subtracted from each pair-wise difference in (7), but in R_β we cannot pull out the pair-wise effects β_m, $m = 1, 2, \ldots, M$, and summarize them by a single number as in $D_{\bar{\beta}}$. Thus, one cannot generalize (7) and (8) to the case of heterogeneous treatment effects.

10. APPENDIX: SENSITIVITY ANALYSIS

Given $p^+ \equiv \frac{\Gamma}{1+\Gamma} \geq 0.5$ and $p^- \equiv \frac{1}{1+\Gamma} \leq 0.5$, define R^+ (R^-) as the sum of M-many independent random variables where the mth variable takes the value r_m with probability p^+ (p^-) and 0 with probability $1-p^+$ ($1-p^-$). Writing R^+ as $\sum_{m=1}^{M} r_m u_m$, where $P(u_m = 1) = p^+$ and $P(u_m = 0) = 1 - p^+$, we get

$$\begin{aligned} E(R^+) &= \sum_{m=1}^{M} r_m E(u_m) = p^+ \sum_{m=1}^{M} r_m = \frac{p^+ M(M+1)}{2}, \\ V(R^+) &= \sum_{m=1}^{M} r_m^2 V(u_m) = p^+(1-p^+) \sum_{m=1}^{M} r_m^2 = \frac{p^+(1-p^+)M(M+1)(2M+1)}{6}. \end{aligned}$$

Analogously, writing R^- as $\sum_{m=1}^{M} r_m u_m$, where $P(u_m = 1) = p^-$ and $P(u_m = 0) = 1 - p^-$, we obtain

$$E(R^-) = \frac{p^- M(M+1)}{2} \quad \text{and} \quad V(R^-) = \frac{p^-(1-p^-)M(M+1)(2M+1)}{6}.$$

It follows from Rosenbaum (2002, Proposition 13) that $P(R^+ \geq a) \geq P(R' \geq a) \geq P(R^- \geq a)$ for arbitrary a.

For treatment effect analysis with matching, various sensitivity analyses have appeared in the statistics literature as reviewed in Rosenbaum (2002), but not many in econometrics. Those that have appeared in the econometrics literature include the parametric/structural regression approach of Imbens (2003) and Altonji et al. (2005). This approach allows for an unobserved confounder to affect both treatment and response, but is heavily dependent on the parametric assumptions about the structural equations of treatment and response.

Ichino et al. (2008) suggested an alternative, simulation-based, approach of sensitivity analysis for matching estimators. This approach also allows for an unobserved confounder to affect both treatment and response, but without relying on any parametric/structural model for the treatment and response. The unobserved confounder is simulated and included in the list of matching covariates to evaluate the sensitivity of point effect estimates. This is feasible only for binary unobserved confounders in the context of binary treatment/response variables, so that the distribution of the unobserved confounder can be characterized by four probability parameters conditional on the treatment/response outcomes.

Gastwirth et al. (1998) extended the Rosenbaum (2002) approach by allowing the unobserved confounder to affect both treatment and response. The approach of Gastwirth et al. (1998) is, however, parametric/structural; it specifies exactly how the unobserved confounder appears in the treatment and response equations. For instance, in the case where both the treatment and response variables are binary, the logit form is obtained, which may not look so objectionable; in other cases, the parametric specification becomes too restrictive. In a sense, the benefit of considering how the unobserved confounder affects the response is obtained at this parametrization cost. Refer to Lee et al. (2007) and Lee and Lee (2009) for applications of this approach. Since a hidden bias results from unobserved confounders affecting both treatment and response, the Rosenbaum (2002) analysis is conservative in the sense that it may be concerned with a hidden bias that does not exist at all if the unobserved confounder does not affect the response. Thus, if we find a result to be robust using the Rosenbaum (2002) approach, its robustness using the Gastwirth et al. (1998) approach is implied. Refer also to Lee (2004) for a nonparametric reduced-form sensitivity analysis.

REFERENCES

Aakvik, A., 2001. Bounding a matching estimator: The case of a Norwegian training program. Oxford Bulletin of Economics and Statistics 63 (1), 115–143.

Abadie, A., Imbens, G. W., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74 (1), 235–267.

Altonji, J. G., Elder, T. E., Taber, C. R., 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. Journal of Political Economy 113 (1), 151–184.

Anderson, J. E., 1979. A theoretical foundation for the gravity equation. American Economic Review 69 (1), 106–116.

Anderson, J. E., van Wincoop, E., 2003. Gravity with gravitas: A solution to the border puzzle. American Economic Review 93 (1), 170–192.

Bagwell, K., Staiger, R. W., 1999. An economic theory of GATT. American Economic Review 89 (1), 215–248.

Bagwell, K., Staiger, R. W., 2001. Reciprocity, non-discrimination and preferential agreements in the multilateral trading system. European Journal of Political Economy 17 (2), 281–325.

Bagwell, K., Staiger, R. W., 2010. The world trade organization: Theory and practice. Annual Review of Economics 2, 223–256.

Baier, S. L., Bergstrand, J. H., 2007. Do free trade agreements actually increase members' international trade? Journal of International Economics 71 (1), 72–95.

Baier, S. L., Bergstrand, J. H., 2009a. Bonus vetus OLS: A simple method for approximating international trade-cost effects using the gravity equation. Journal of International Economics 77 (1), 77–85.

Baier, S. L., Bergstrand, J. H., 2009b. Estimating the effects of free trade agreements on trade flows using matching econometrics. Journal of International Economics 77 (1), 63–76.

Bergstrand, J. H., 1985. The gravity equation in international trade: Some microeconomic foundations and empirical evidence. Review of Economics and Statistics 67 (3), 474–481.

Broda, C., Limão, N., Weinstein, D. E., 2008. Optimal tariffs and market power: The evidence. American Economic Review 98 (5), 2032–2065.

Caliendo, M., Hujer, R., Thomsen, S. L., 2005. The employment effects of job creation schemes in Germany: A microeconometric evaluation. IZA Discussion Paper 1512.

Chang, P.-L., Lee, M.-J., 2010. The WTO trade effect, SMU Economics and Statistics Working Paper No. 31–2010.

Deardorff, A. V., 1998. Determinants of bilateral trade: Does gravity work in a neoclassical world? In: Frankel, J. A. (Ed.), The Regionalization of the World Economy. University of Chicago Press, Chicago, pp. 7–22.

Ernst, M. D., 2004. Permutation methods: A basis for exact inference. Statistical Science 19 (4), 676–685.

Felbermayr, G., Kohler, W., 2007. Does WTO membership make a difference at the extensive margin of world trade? CESifo Working Paper 1898.

Fisher, R. A., 1935. The Design of Experiments. Oliver and Boyd, London.

Frankel, J. A., 1997. Regional Trading Blocs in the World Economic System. Institute for International Economics, Washington, DC.

Gastwirth, J. L., Krieger, A. M., Rosenbaum, P. R., 1998. Dual and simultaneous sensitivity analysis for matched pairs. Biometrika 85 (4), 907–920.

Heckman, J. J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Heckman, J. J., Ichimura, H., Smith, J., Todd, P. E., 1998. Characterizing selection bias using experimental data. Econometrica 66 (5), 1017–1098.

Heckman, J. J., Ichimura, H., Todd, P. E., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 64 (4), 605–654.

Helpman, E., Melitz, M., Rubinstein, Y., 2008. Estimating trade flows: Trading partners and trading volumes. Quarterly Journal of Economics 123 (2), 441–487.

Ho, D. E., Imai, K., 2006. Randomization inference with natural experiments: an analysis of ballot effects in the 2003 California recall election. Journal of the American Statistical Association 101, 888–900.

Hodges, J., Lehmann, E., 1963. Estimates of location based on rank tests. Annals of Mathematical Statistics 34 (2), 598–611.

Hollander, M., Wolfe, D. A., 1999. Nonparametric Statistical Methods, 2nd Edition. Wiley, New York.

Hujer, R., Caliendo, M., Thomsen, S. L., 2004. New evidence on the effects of job creation schemes in Germany – a matching approach with threefold heterogeneity. Research in Economics 58 (4), 257–302.

Hujer, R., Thomsen, S. L., 2006. How do employment effects of job creation schemes differ with respect to the foregoing unemployment duration? ZEW Discussion Paper 06–047.

Ichino, A., Mealli, F., Nannicini, T., 2008. From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? Journal of Applied Econometrics 23 (3), 305–327.

Imbens, G. W., 2003. Sensitivity to exogeneity assumptions in program evaluation. American Economic Review (Papers and Proceedings) 93 (2), 126–132.

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

Imbens, G. W., Rosenbaum, P. R., 2005. Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. Journal of the Royal Statistical Society Ser. A, 168 (1), 109–126.

Imbens, G. W., Wooldridge, J. M., 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47, 5–86.

Johnson, H. G., 1953–1954. Optimum tariffs and retaliation. Review of Economic Studies 21 (2), 142–153.

Lechner, M., 2000. An evaluation of public-sector-sponsored continuous vocational training programs in East Germany. Journal of Human Resources 35 (2), 347–375.

Lee, M.-J., 2005. Micro-econometrics for Policy, Program, and Treatment Effects. Oxford University Press, New York.

Lee, M.-J., 2010. Treatment effects in sample selection models and their nonparametric estimation, manuscript.

Lee, M.-J., Häkkinen, U., Rosenqvist, G., 2007. Finding the best treatment under heavy censoring and hidden bias. Journal of the Royal Statistical Society Ser. A, 170 (1), 133–147.

Lee, M.-J., Lee, S.-J., 2009. Sensitivity analysis of job-training effects on reemployment for Korean women. Empirical Economics 36 (1), 81–107.

Lehmann, E. L., Romano, J. P., 2005. Testing Statistical Hypotheses, 3rd Edition. Springer.

Lu, B., Zanutto, E., Hornik, R., Rosenbaum, P. R., 2001. Matching with doses in an observational study of a media campaign against drug abuse. Journal of the American Statistical Association 96 (456), 1245–1253.

Persson, T., 2001. Currency unions and trade: How large is the treatment effect? Economic Policy 16 (33), 435–448.

Pesarin, F., 2001. Multivariate Permutation Tests: With Applications in Biostatistics. Wiley, Chichester.

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

Rose, A. K., 2004. Do we really know that the WTO increases trade? American Economic Review 94 (1), 98–114.

Rose, A. K., 2010. The effect of membership in the GATT/WTO on trade: Where do we stand? In: Drabek, Z. (Ed.), Is the World Trade Organization Attractive Enough For Emerging Economies? Palgrave Macmillan.

Rosenbaum, P. R., 2002. Observational Studies, 2nd Edition. Springer.

Staiger, R. W., Tabellini, G., 1987. Discretionary trade policy and excessive protection. American Economic Review 77 (5), 823–837.

Staiger, R. W., Tabellini, G., 1989. Rules and discretion in trade policy. European Economic Review 33 (6), 1265–1277.

Staiger, R. W., Tabellini, G., 1999. Do GATT rules help governments make domestic commitments? Economics & Politics 11 (2), 109–144.

Subramanian, A., Wei, S.-J., 2007. The WTO promotes trade, strongly but unevenly. Journal of International Economics 72 (1), 151–175.

Tomz, M., Goldstein, J., Rivers, D., 2007. Do we really know that the WTO increases trade? comment. American Economic Review 97 (5), 2005–2018.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1 (6), 80–83.

Table 1: Variables and definitions

Variable	Definition
response variable:	
ltrade	the log average value of a dyad's current real bilateral trade flows
covariates:	
ldist	the log distance between the two countries in a dyad
lrgdp	the log product of a dyad's real GDPs
lrgdppc	the log product of a dyad's real GDPs per capita
comlang	= 1 if the two countries in a dyad share a common language (= 0 otherwise)
border	= 1 if the two countries in a dyad share a land border (= 0 otherwise)
landl	= the number of landlocked countries in a dyad
island	= the number of island nations in a dyad
lareap	the log product of a dyad's land areas
comcol	= 1 if the two countries in a dyad were ever colonies after 1945 with the same colonizer (= 0 otherwise)
curcol	= 1 if the two countries in a dyad are in a colonial relationship (= 0 otherwise)
colony	= 1 if the two countries in a dyad were ever in a colonial relationship (= 0 otherwise)
comctry	= 1 if the two countries in a dyad remained part of the same nation during the sample period (= 0 otherwise)
custrict	= 1 if the two countries in a dyad use the same currency (= 0 otherwise)
regional	= 1 if the two countries in a dyad belong to the same regional trade agreement (= 0 otherwise)
year dummy	for $t = 1948, \ldots, 1999$.
treatment variables:	(the variable becomes part of the covariates if not used as a treatment variable)
bothin	= 1 if both countries in a dyad are GATT/WTO members (= 0 otherwise)
onein	= 1 if only one country in a dyad is a GATT/WTO member (= 0 otherwise)
gsp	= 1 if the two countries in a dyad have a GSP arrangement (= 0 otherwise)

Table 2: Rose (2004) data set – descriptive statistics

variables	Both in				One in				None in (control group)			
	mean	SD	25%	75%	mean	SD	25%	75%	mean	SD	25%	75%
ltrade	10.472	3.415	8.344	12.815	9.759	3.253	8.013	11.937	9.246	2.964	8.062	11.124
ldist	8.198	0.797	7.843	8.745	8.188	0.772	7.751	8.749	7.873	0.972	7.216	8.685
lrgdp	48.404	2.681	46.615	50.218	47.582	2.526	45.930	49.265	46.432	2.582	44.968	48.068
lrgdppc	16.234	1.579	15.242	17.358	15.940	1.394	15.036	16.902	15.386	1.344	14.508	16.249
comlang	0.238	0.426	0	0	0.187	0.390	0	0	0.304	0.460	0	1
border	0.027	0.162	0	0	0.026	0.160	0	0	0.072	0.258	0	0
landl	0.251	0.471	0	0	0.241	0.461	0	0	0.246	0.467	0	0
island	0.364	0.548	0	1	0.331	0.535	0	1	0.264	0.503	0	0
lareap	24.145	3.230	22.445	26.314	24.270	3.293	22.466	26.588	24.238	3.480	22.362	26.739
comcol	0.105	0.307	0	0	0.089	0.285	0	0	0.124	0.330	0	0
curcol	0.004	0.062	0	0	0	0	0	0	0.000	0.017	0	0
colony	0.027	0.162	0	0	0.016	0.126	0	0	0.008	0.092	0	0
comctry	0.001	0.024	0	0	0	0	0	0	0	0	0	0
custrict	0.019	0.136	0	0	0.009	0.093	0	0	0.014	0.117	0	0
regional	0.018	0.134	0	0	0.009	0.096	0	0	0.019	0.138	0	0
gsp	0.299	0.458	0	1	0.201	0.400	0	0	0.008	0.090	0	0
year	1984.1	11.5	1976	1994	1978.9	12.4	1970	1989	1973.6	12.7	1963	1983
obs.	114,750				98,810				21,037			

Table 3: Rose (2004) data set – selection on observables

variables	Both in				One in			
	odds	p-value	95% CI		odds	p-value	95% CI	
ldist	1.174	0.000	1.147	1.202	1.230	0.000	1.203	1.257
lrgdp	1.538	0.000	1.521	1.555	1.222	0.000	1.209	1.235
lrgdppc	0.892	0.000	0.878	0.906	0.999	0.907	0.984	1.014
comlang	0.767	0.000	0.735	0.801	0.714	0.000	0.686	0.743
border	0.870	0.002	0.795	0.951	0.848	0.000	0.783	0.918
landl	1.187	0.000	1.142	1.233	1.072	0.000	1.034	1.112
island	1.872	0.000	1.793	1.955	1.448	0.000	1.391	1.508
lareap	0.875	0.000	0.867	0.882	0.947	0.000	0.940	0.955
comcol	1.645	0.000	1.546	1.750	1.293	0.000	1.223	1.368
curcol	12.385	0.000	5.320	28.834	1.678	0.000	1.417	1.988
colony	2.126	0.000	1.775	2.547	—	—	—	—
comctry	—	—	—	—	—	—	—	—
custrict	6.961	0.000	6.031	8.034	1.705	0.000	1.467	1.981
regional	0.762	0.000	0.666	0.873	0.879	0.070	0.764	1.011
gsp	27.698	0.000	23.750	32.303	19.230	0.000	16.487	22.428
obs.	135,720				119,841			

Note: The results are based on logistic regressions with *nonein* = 1 observations as the control group. The odds estimates are equal to exponential transformation of coefficient estimates in logit regressions. All regressions include year dummies. In the both-in regression, *comctry* is dropped as *comctry* = 1 predicts *bothin* = 1 perfectly. In the one-in regression, *curcol* is dropped as *curcol* = 1 predicts *onein* = 0 perfectly and *comctry* is dropped because of collinearity.

Figure 1: Support of covariates for the treatment and control groups

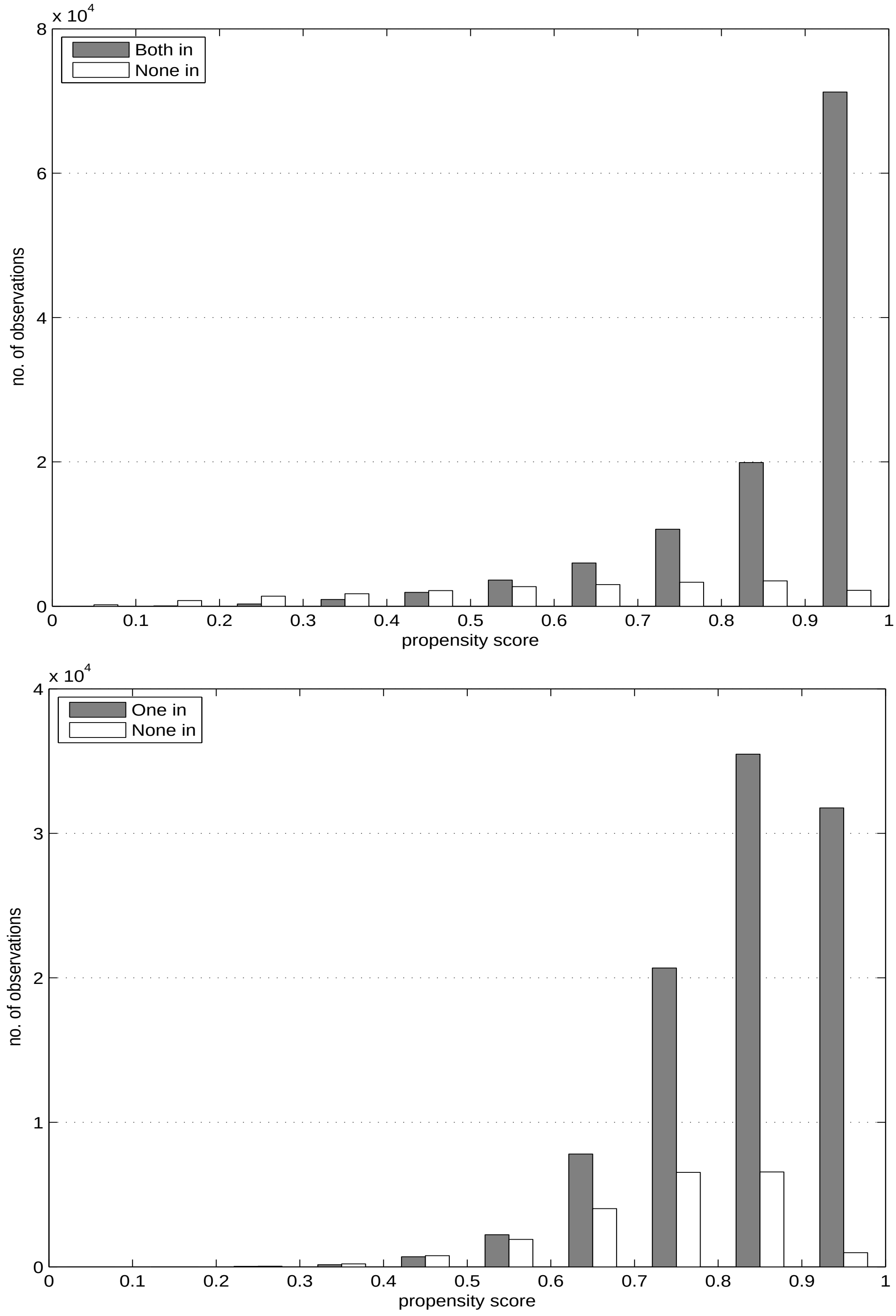


Table 4: Rose (2004) data set – unrestricted matching

	permutation test			signed-rank test			sensitivity analysis			
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	one-sided test		two-sided test	
caliper	effect	p-value	95% CI	effect	p-value	95% CI	Γ^*	as in	Γ^*	as in
Both in GATT/WTO treatment effect										
on the treated ($M_1 = 114,750$)**:**										
100%	1.328	0.000	[1.307, 1.349]	1.332	0.000	[1.312, 1.351]	2.434	R^+	2.428	R^+
80%	1.075	0.000	[1.052, 1.098]	1.075	0.000	[1.053, 1.096]	2.086	R^+	2.081	R^+
60%	0.836	0.000	[0.810, 0.862]	0.835	0.000	[0.810, 0.859]	1.780	R^+	1.775	R^+
40%	0.553	0.000	[0.522, 0.584]	0.535	0.000	[0.507, 0.563]	1.472	R^+	1.467	R^+
on the untreated ($M_0 = 21,037$)**:**										
100%	0.337	0.000	[0.296, 0.379]	0.303	0.000	[0.266, 0.342]	1.250	R^+	1.243	R^+
80%	0.239	0.000	[0.192, 0.286]	0.200	0.000	[0.157, 0.241]	1.144	R^+	1.138	R^+
60%	0.185	0.000	[0.131, 0.239]	0.138	0.000	[0.090, 0.187]	1.084	R^+	1.077	R^+
40%	0.304	0.000	[0.239, 0.368]	0.243	0.000	[0.184, 0.301]	1.177	R^+	1.167	R^+
on all ($M_1 + M_0 = 135,787$)**:**										
100%	1.175	0.000	[1.156, 1.193]	1.161	0.000	[1.143, 1.179]	2.209	R^+	2.205	R^+
80%	0.899	0.000	[0.878, 0.919]	0.883	0.000	[0.863, 0.902]	1.858	R^+	1.854	R^+
60%	0.636	0.000	[0.613, 0.659]	0.619	0.000	[0.597, 0.640]	1.559	R^+	1.555	R^+
40%	0.428	0.000	[0.400, 0.455]	0.399	0.000	[0.374, 0.424]	1.342	R^+	1.338	R^+
One in GATT/WTO treatment effect										
on the treated ($M_1 = 98,810$)**:**										
100%	0.767	0.000	[0.746, 0.789]	0.773	0.000	[0.753, 0.792]	1.759	R^+	1.755	R^+
80%	0.564	0.000	[0.540, 0.588]	0.568	0.000	[0.547, 0.589]	1.525	R^+	1.521	R^+
60%	0.422	0.000	[0.396, 0.449]	0.428	0.000	[0.405, 0.451]	1.397	R^+	1.393	R^+
40%	0.326	0.000	[0.296, 0.357]	0.325	0.000	[0.298, 0.351]	1.294	R^+	1.289	R^+
on the untreated ($M_0 = 21,037$)**:**										
100%	0.030	0.068	[-0.009, 0.069]	0.034	0.022	[0.000, 0.068]	1.006	R^+	1.001	R^+
80%	0.092	0.000	[0.048, 0.135]	0.089	0.000	[0.052, 0.126]	1.057	R^+	1.051	R^+
60%	0.078	0.001	[0.028, 0.129]	0.084	0.000	[0.041, 0.127]	1.046	R^+	1.039	R^+
40%	0.138	0.000	[0.076, 0.201]	0.149	0.000	[0.096, 0.203]	1.102	R^+	1.094	R^+
on all ($M_1 + M_0 = 119,847$)**:**										
100%	0.638	0.000	[0.619, 0.657]	0.632	0.000	[0.615, 0.649]	1.610	R^+	1.607	R^+
80%	0.443	0.000	[0.422, 0.464]	0.437	0.000	[0.418, 0.455]	1.401	R^+	1.397	R^+
60%	0.324	0.000	[0.301, 0.347]	0.321	0.000	[0.301, 0.340]	1.297	R^+	1.293	R^+
40%	0.225	0.000	[0.198, 0.253]	0.220	0.000	[0.197, 0.243]	1.194	R^+	1.190	R^+
GSP treatment effect										
on the treated ($M_1 = 54,285$)**:**										
100%	0.851	0.000	[0.831, 0.871]	0.792	0.000	[0.774, 0.811]	2.277	R^+	2.269	R^+
80%	0.757	0.000	[0.736, 0.778]	0.696	0.000	[0.676, 0.716]	2.125	R^+	2.117	R^+
60%	0.693	0.000	[0.668, 0.717]	0.627	0.000	[0.604, 0.649]	1.998	R^+	1.990	R^+
40%	0.665	0.000	[0.635, 0.696]	0.581	0.000	[0.553, 0.608]	1.879	R^+	1.869	R^+

Note:
1. The pool of potential matches for an observation is restricted to observations with the opposite treatment status; no further restriction is imposed. The number of matched pairs for the effect on the treated (untreated) is indicated by M_1 (M_0).
2. The caliper is set such that only the best 100%, 80%, 60%, or 40% of matched pairs obtained are included in the analysis. For example, with the caliper choice of 60%, the matched pairs with the scale-normalized distance exceeding the upper 60th percentile of all matched pairs obtained are discarded.
3. In 'permutation test', the results are based on the D-statistic.
4. In 'signed-rank test', the results are based on the R-statistic.
5. We carried out both simulation and normal approximation approaches for calculating the p-values and the CI's, and found almost identical results (which is expected given that the sample size is large). Thus, we report only the results based on normal approximation.
6. In 'sensitivity analysis', the sensitivity analysis is conducted for the significance (p-value) of the signed-rank R-statistic based on the critical level $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of the odds ratio Γ) indicates the relevant distribution in calculating the critical bound Γ^* at which the conclusion of the signed-rank test reverses.

Table 5: Rose (2004) data set – restricted matching effect estimates and sensitivity

	unrestricted		within dyad		within year		within devel.	
caliper	effect	Γ^*	effect	Γ^*	effect	Γ^*	effect	Γ^*
			Both in GATT/WTO treatment effect					
on the treated:								
M_1	114,750		19,760		114,750		112,959	
100%	1.328***	2.428	0.941***	3.170	1.329***	2.427	1.124***	2.019
80%	1.075***	2.081	0.760***	2.543	1.075***	2.081	0.778***	1.601
60%	0.836***	1.775	0.833***	2.771	0.836***	1.775	0.541***	1.385
40%	0.553***	1.467	0.796***	2.503	0.553***	1.467	0.393***	1.256
			One in GATT/WTO treatment effect					
on the treated:								
M_1	98,810		23,463		98,810		98,363	
100%	0.767***	1.755	0.464***	1.931	0.761***	1.747	0.650***	1.552
80%	0.564***	1.521	0.403***	1.772	0.564***	1.521	0.476***	1.391
60%	0.422***	1.393	0.371***	1.656	0.422***	1.393	0.342***	1.263
40%	0.326***	1.289	0.314***	1.508	0.326***	1.289	0.242***	1.197
			GSP treatment effect					
on the treated:								
M_1	54,285		52,025		54,285		53,811	
100%	0.851***	2.269	0.487***	2.570	0.850***	2.267	0.732***	2.011
80%	0.757***	2.117	0.492***	2.494	0.757***	2.117	0.588***	1.807
60%	0.693***	1.990	0.379***	1.937	0.693***	1.990	0.507***	1.699
40%	0.665***	1.869	0.271***	1.528	0.665***	1.869	0.410***	1.530

Note:
The effect estimate refers to the D-statistic. All significance levels refer to a two-sided test. The effect estimate is significant at the 1%, 5%, or 10% significance level if indicated by a superscript of ***, **, or *, respectively. The sensitivity parameter Γ^* is based on a two-sided test at the 5% significance level. The distribution used in calculating the critical bound Γ^* is R^+ unless a superscript $^-$ is indicated following the bound Γ^*, in which case, R^- is used. Unless otherwise indicated, the significance level of the D-statistic agrees with that of the R-statistic.

Table 6: Tomz et al. (2007) data set – matching effect estimates and sensitivity

	unrestricted		within dyad		within year		within devel.	
caliper	effect	Γ^*	effect	Γ^*	effect	Γ^*	effect	Γ^*
			Both participating in GATT/WTO treatment effect					
on the treated:								
M_1	152,986		8,005		152,986		152,986	
100%	1.418***	2.426	1.554***	7.535	1.427***	2.439	1.065***	2.099
80%	1.260***	2.284	1.513***	6.689	1.260***	2.284	0.710***	1.626
60%	1.089***	2.058	1.285***	4.969	1.089***	2.058	0.515***	1.382
40%	0.762***	1.706	1.361***	5.134	0.762***	1.706	0.461***	1.324
			One participating in GATT/WTO treatment effect					
on the treated:								
M_1	71,908		11,637		71,908		71,908	
100%	0.818***	1.777	0.852***	2.877	0.822***	1.782	0.464***	1.457
80%	0.631***	1.580	0.716***	2.393	0.631***	1.580	0.278***	1.244
60%	0.444***	1.423	0.738***	2.279	0.444***	1.423	0.290***	1.241
40%	0.304***	1.295	0.546***	1.840	0.304***	1.295	0.172***	1.154
			GSP treatment effect					
on the treated:								
M_1	54,285		52,025		54,285		54,285	
100%	0.824***	2.243	0.485***	2.561	0.823***	2.241	0.688***	1.959
80%	0.726***	2.065	0.480***	2.407	0.726***	2.065	0.569***	1.786
60%	0.667***	1.944	0.375***	1.893	0.667***	1.944	0.489***	1.679
40%	0.621***	1.782	0.265***	1.494	0.621***	1.782	0.401***	1.510

Note: The general notes for Table 5 apply to the current table.

Table 7: Rose (2004) data set – kernel-weighting matching effect estimates

caliper	unrestricted	within dyad	within year	within devel.
	Both in GATT/WTO treatment effect			
on the treated:				
100%	1.323	0.929	1.284	0.962
80%	1.078	0.764	1.076	0.778
60%	0.840	0.835	0.837	0.542
40%	0.558	0.799	0.554	0.396
	One in GATT/WTO treatment effect			
on the treated:				
100%	0.753	0.484	0.748	0.604
80%	0.573	0.423	0.571	0.483
60%	0.436	0.393	0.433	0.353
40%	0.344	0.342	0.341	0.260
	GSP treatment effect			
on the treated:				
100%	0.874	0.491	0.863	0.744
80%	0.786	0.497	0.773	0.605
60%	0.731	0.384	0.712	0.544
40%	0.709	0.277	0.688	0.456

Table 8: Rose (2004) data set – trading relationship exists before GATT/WTO membership

	unrestricted		within dyad		within year		within devel.	
caliper	effect	Γ^*	effect	Γ^*	effect	Γ^*	effect	Γ^*
			Both in GATT/WTO treatment effect					
on the treated:								
M_1	19,760		19,760		19,760		19,522	
100%	1.599***	2.983	1.032***	3.372	1.606***	2.983	1.302***	2.364
80%	1.447***	2.660	0.836***	2.726	1.447***	2.660	1.157***	2.086
60%	1.149***	2.195	0.886***	2.885	1.149***	2.195	0.909***	1.771
40%	0.861***	1.817	0.821***	2.586	0.861***	1.817	0.639***	1.469
			One in GATT/WTO treatment effect					
on the treated:								
M_1	23,463		23,463		23,463		23,384	
100%	0.986***	2.060	0.469***	1.935	0.985***	2.058	0.903***	1.879
80%	0.758***	1.743	0.392***	1.753	0.758***	1.743	0.691***	1.609
60%	0.615***	1.590	0.351***	1.653	0.615***	1.590	0.492***	1.384
40%	0.535***	1.491	0.354***	1.621	0.535***	1.491	0.415***	1.320

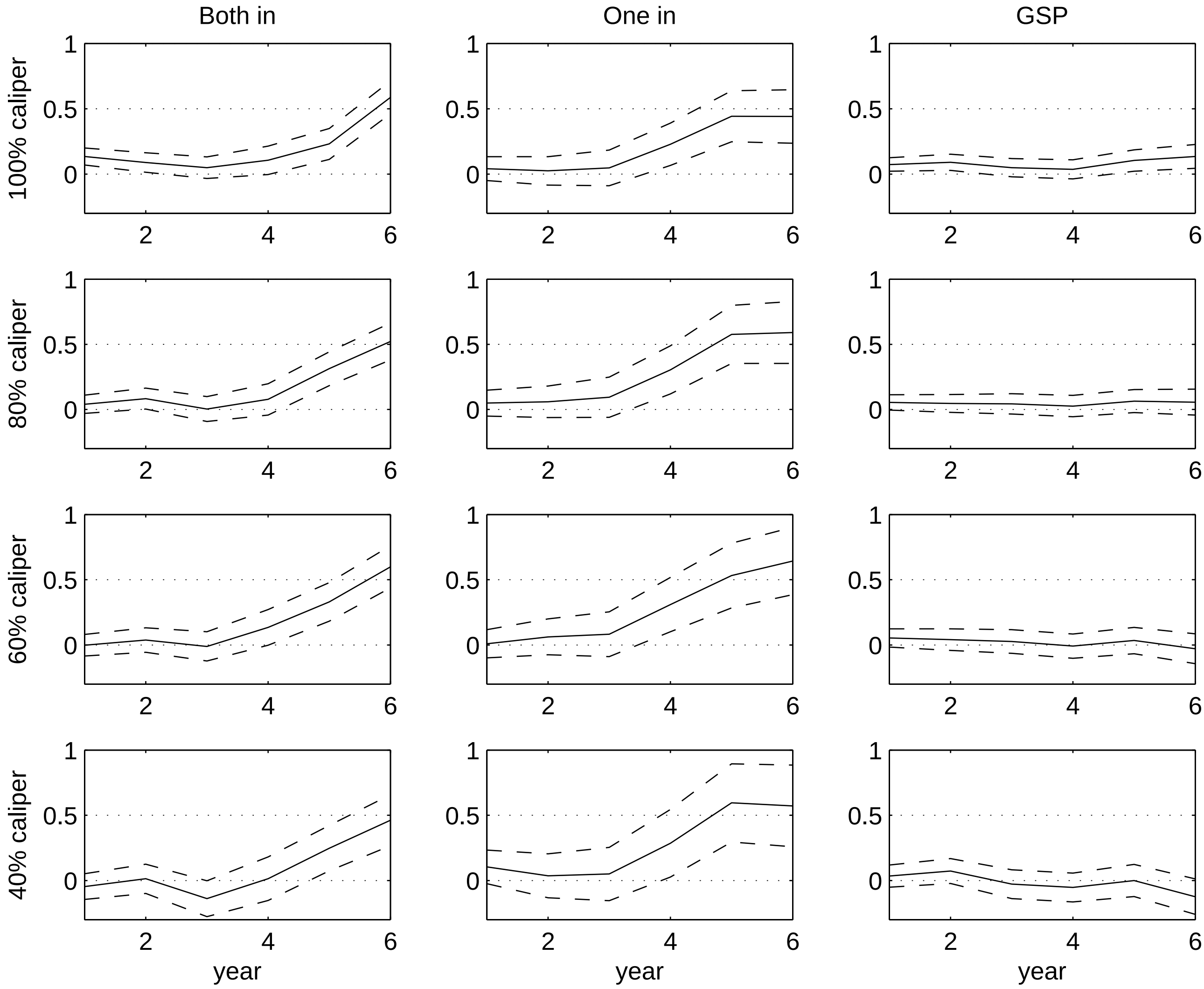
Note: The general notes for Table 5 apply to the current table.

Table 9: Rose (2004) data set – with multilateral resistance terms

	unrestricted		within dyad		within year		within devel.	
caliper	effect	Γ^*	effect	Γ^*	effect	Γ^*	effect	Γ^*
	Both in GATT/WTO treatment effect							
on the treated:								
M_1	114,750		19,760		114,750		112,959	
100%	1.622***	2.616	0.942***	3.170	1.618***	2.605	1.243***	2.041
80%	1.355***	2.273	0.778***	2.594	1.355***	2.273	0.750***	1.543
60%	1.130***	2.023	0.850***	2.858	1.130***	2.023	0.659***	1.452
40%	0.894***	1.798	0.845***	2.624	0.894***	1.798	0.569***	1.375
	One in GATT/WTO treatment effect							
on the treated:								
M_1	98,810		23,463		98,810		98,363	
100%	0.627***	1.560	0.454***	1.903	0.627***	1.560	0.455***	1.385
80%	0.401***	1.368	0.399***	1.761	0.401***	1.368	0.270***	1.230
60%	0.246***	1.242	0.371***	1.650	0.246***	1.242	0.209***	1.194
40%	0.252***	1.267	0.374***	1.612	0.252***	1.267	0.107***	1.124
	GSP treatment effect							
on the treated:								
M_1	54,285		52,025		54,285		53,811	
100%	1.044***	2.243	0.485***	2.559	1.043***	2.242	0.954***	2.183
80%	1.060***	2.309	0.494***	2.624	1.060***	2.309	0.948***	2.195
60%	0.954***	2.139	0.456***	2.261	0.954***	2.139	0.762***	1.869
40%	0.872***	2.023	0.325***	1.679	0.872***	2.023	0.712***	1.748

Note: The general notes for Table 5 apply to the current table.

Figure 2: Difference-in-Difference matching estimates



Note:
1. The horizontal axis indicates the years of lead and lag (a, b) used in the DD estimation; here, symmetric leads and lags are used. The vertical axis (not labeled) indicates the treatment effect magnitude.
2. The solid line indicates the treatment effect point estimate. The dashed lines indicate the 95% CI based on the permutation test.
3. The sample size (the number of qualified matched pairs) for each treatment scenario is as follows.
Both-in: 3600 (1 year), 3216 (2 years), 2955 (3 years), 2461 (4 years), 2277 (5 years), 1812 (6 years).
One-in: 1303 (1 year), 1110 (2 years), 1022 (3 years), 828 (4 years), 736 (5 years), 651 (6 years).
GSP: 2231 (1 year), 2184 (2 years), 2031 (3 years), 1976 (4 years), 1913 (5 years), 1859 (6 years).
These correspond to the sample size used in the 100% caliper choice.

Table 10: Rose (2004) data set – placebo exercise

		years before the actual treatment year					
		12		11		10	
DD window (years)	caliper	effect	95% CI	effect	95% CI	effect	95% CI
1	100%	0.040	[-0.107, 0.187]	-0.064	[-0.219, 0.090]	0.052	[-0.088, 0.192]
	80%	-0.003	[-0.171, 0.166]	-0.118	[-0.289, 0.053]	0.037	[-0.115, 0.189]
	60%	-0.025	[-0.224, 0.174]	-0.202	*[-0.390, -0.013]*	0.026	[-0.156, 0.207]
	40%	-0.024	[-0.243, 0.195]	-0.264	*[-0.498, -0.030]*	0.014	[-0.209, 0.238]
2	100%	-0.025	[-0.201, 0.152]	0.010	[-0.161, 0.180]	-0.006	[-0.180, 0.169]
	80%	-0.080	[-0.282, 0.123]	-0.088	[-0.283, 0.106]	-0.006	[-0.202, 0.191]
	60%	-0.038	[-0.276, 0.199]	-0.044	[-0.279, 0.191]	-0.061	[-0.290, 0.167]
	40%	-0.054	[-0.360, 0.253]	-0.050	[-0.337, 0.237]	-0.065	[-0.340, 0.209]
3	100%	0.092	[-0.139, 0.324]	0.145	[-0.057, 0.346]	0.010	[-0.175, 0.196]
	80%	0.196	[-0.070, 0.463]	0.181	[-0.052, 0.415]	-0.044	[-0.249, 0.160]
	60%	0.234	[-0.039, 0.507]	0.067	[-0.182, 0.317]	-0.065	[-0.294, 0.164]
	40%	0.061	[-0.261, 0.383]	0.146	[-0.148, 0.440]	-0.089	[-0.353, 0.175]
4	100%	0.012	[-0.198, 0.222]	-0.027	[-0.239, 0.185]	-0.019	[-0.223, 0.185]
	80%	0.024	[-0.208, 0.257]	0.057	[-0.196, 0.311]	0.039	[-0.190, 0.268]
	60%	-0.007	[-0.277, 0.263]	0.007	[-0.294, 0.308]	0.122	[-0.137, 0.382]
	40%	-0.007	[-0.363, 0.349]	-0.062	[-0.426, 0.301]	0.164	[-0.151, 0.479]
5	100%	-0.166	[-0.411, 0.079]	-0.213	[-0.438, 0.012]		
	80%	-0.116	[-0.391, 0.159]	-0.240	[-0.491, 0.011]		
	60%	-0.121	[-0.434, 0.193]	-0.304	*[-0.606, -0.003]*		
	40%	-0.133	[-0.508, 0.243]	-0.304	[-0.658, 0.049]		
6	100%	-0.138	[-0.425, 0.149]				
	80%	-0.105	[-0.430, 0.220]				
	60%	-0.015	[-0.388, 0.357]				
	40%	-0.230	[-0.651, 0.191]				
		years before the actual treatment year					
		9		8		7	
		effect	95% CI	effect	95% CI	effect	95% CI
1	100%	0.134	[-0.011, 0.278]	-0.039	[-0.172, 0.093]	0.106	[-0.031, 0.243]
	80%	0.090	[-0.065, 0.244]	-0.035	[-0.169, 0.100]	0.078	[-0.069, 0.225]
	60%	0.061	[-0.115, 0.238]	-0.002	[-0.164, 0.160]	0.036	[-0.135, 0.206]
	40%	0.116	[-0.091, 0.324]	-0.011	[-0.214, 0.192]	0.021	[-0.187, 0.230]
2	100%	-0.072	[-0.232, 0.089]	0.058	[-0.094, 0.209]		
	80%	-0.072	[-0.255, 0.111]	0.058	[-0.097, 0.214]		
	60%	-0.047	[-0.245, 0.150]	-0.004	[-0.174, 0.167]		
	40%	-0.009	[-0.264, 0.246]	-0.003	[-0.213, 0.206]		
3	100%	-0.104	[-0.280, 0.073]				
	80%	-0.077	[-0.271, 0.117]				
	60%	-0.044	[-0.263, 0.175]				
	40%	-0.084	[-0.359, 0.191]				

Note:
1. The estimation proceeds as described in Section 5.6 for DD estimation, but with a bogus treatment year $t' = t - d$ used, where $\{d = 7, \dots, 12\}$, which predates the actual year of treatment t (here identified as the first year when either country in a treated dyad joins the GATT/WTO).
2. The DD window refers to the years of lead and lag (a, b) used in the DD estimation, where it is set that $a = b$.
3. The effect refers to the bogus treatment effect on the treated dyad when using the bogus treatment year.
4. The effect estimates that are significantly negative are indicated by CI's in italics.

Table 11: parametric gravity estimates with heterogeneous treatment effects

ltrade	Rose default		heter. both-in effect		heter. both-in / one-in effect		heter. both-in / one-in / gsp effect	
ldist	-1.119	(0.022)	-1.112	(0.028)	-1.099	(0.060)	-1.100	(0.060)
lrgdp	0.916	(0.010)	0.900	(0.012)	0.858	(0.027)	0.858	(0.027)
lrgdppc	0.321	(0.014)	0.246	(0.019)	0.045	(0.044)	0.044	(0.044)
comlang	0.313	(0.040)	0.259	(0.053)	0.092	(0.107)	0.091	(0.107)
border	0.526	(0.111)	0.475	(0.122)	0.560	(0.190)	0.558	(0.190)
landl	-0.271	(0.031)	-0.253	(0.041)	-0.174	(0.086)	-0.173	(0.086)
island	0.042	(0.036)	0.043	(0.048)	0.108	(0.116)	0.109	(0.116)
lareap	-0.097	(0.008)	-0.122	(0.010)	-0.171	(0.023)	-0.171	(0.023)
comcol	0.585	(0.067)	0.669	(0.084)	1.080	(0.158)	1.079	(0.158)
curcol	1.075	(0.235)	2.780	(0.356)	4.812	(0.570)	4.810	(0.570)
colony	1.164	(0.117)	1.076	(0.152)	-0.526	(0.210)	-0.522	(0.209)
comctry	-0.016	(1.081)	0.056	(1.035)	0.047	(1.035)	0.333	(1.035)
custrict	1.118	(0.122)	0.624	(0.177)	0.038	(0.325)	0.037	(0.324)
regional	1.199	(0.106)	1.435	(0.154)	0.576	(0.392)	0.573	(0.391)
bothin	-0.042	(0.053)	-4.587	(0.636)	-10.720	(1.102)	-10.260	(1.124)
onein	-0.058	(0.049)	-0.056	(0.048)	-7.606	(1.075)	-7.402	(1.078)
gsp	0.859	(0.032)	1.127	(0.048)	0.556	(0.258)	-2.214	(0.760)
bothin x ldist			-0.017	(0.037)	-0.030	(0.065)	-0.054	(0.066)
bothin x lrgdp			0.029	(0.016)	0.071	(0.029)	0.057	(0.030)
bothin x lrgdppc			0.134	(0.025)	0.335	(0.047)	0.343	(0.048)
bothin x comlang			0.134	(0.067)	0.301	(0.117)	0.248	(0.121)
bothin x border			0.109	(0.197)	0.024	(0.254)	0.027	(0.250)
bothin x landl			-0.048	(0.052)	-0.127	(0.093)	-0.117	(0.096)
bothin x island			-0.035	(0.059)	-0.101	(0.123)	-0.075	(0.124)
bothin x lareap			0.052	(0.013)	0.101	(0.024)	0.114	(0.025)
bothin x comcol			-0.193	(0.114)	-0.606	(0.180)	-0.584	(0.180)
bothin x curcol			-1.890	(0.443)	-3.914	(0.621)	-3.737	(0.637)
bothin x colony			0.088	(0.186)	1.692	(0.253)	1.584	(0.281)
bothin x custrict			0.784	(0.219)	1.374	(0.352)	1.324	(0.352)
bothin x regional			-0.589	(0.193)	0.271	(0.409)	0.237	(0.409)
bothin x gsp			-0.458	(0.054)	0.108	(0.260)	-0.051	(0.281)
onein x ldist					-0.013	(0.063)	-0.030	(0.064)
onein x lrgdp					0.053	(0.029)	0.045	(0.029)
onein x lrgdppc					0.246	(0.047)	0.252	(0.047)
onein x comlang					0.273	(0.116)	0.249	(0.117)
onein x border					-0.077	(0.230)	-0.093	(0.229)
onein x landl					-0.099	(0.091)	-0.094	(0.091)
onein x island					-0.100	(0.119)	-0.083	(0.119)
onein x lareap					0.057	(0.024)	0.065	(0.024)
onein x comcol					-0.580	(0.175)	-0.568	(0.175)
onein x colony					1.708	(0.239)	1.609	(0.261)
onein x custrict					0.674	(0.369)	0.647	(0.374)
onein x regional					1.167	(0.409)	1.079	(0.415)
onein x gsp					0.479	(0.261)	0.362	(0.281)
gsp x ldist							0.180	(0.045)
gsp x lrgdp							0.062	(0.018)
gsp x lrgdppc							-0.018	(0.029)
gsp x comlang							0.188	(0.074)
gsp x border							-1.545	(0.383)
gsp x landl							-0.038	(0.060)
gsp x island							-0.135	(0.064)
gsp x lareap							-0.054	(0.014)
gsp x curcol							-0.585	(0.402)
gsp x colony							0.141	(0.191)
gsp x comctry							-1.421	(1.074)
gsp x custrict							0.169	(0.278)
gsp x regional							0.635	(0.279)
mean bothin effect	-0.042	(0.053)	-0.043	(0.001)	0.272	(0.002)	0.240	(0.002)
mean onein effect	-0.058	(0.049)	-0.056	(0.048)	0.272	(0.002)	0.241	(0.001)
mean gsp effect	0.859	(0.032)	1.127	(0.048)	0.556	(0.258)	0.718	(0.001)
R^2	0.6480		$0.6504^{\dagger}$		$0.6525^{\dagger}$		$0.6530^{\dagger}$	

Note:

1. OLS with year effects (intercepts not reported). Robust standard errors (clustering by dyads) are in the parenthesis. Some interaction terms are dropped due to collinearity.
2. When an effect is heterogeneous, the subject-wise effect equals the main effect plus the interaction effects scaled by the subject's covariates. The mean effect is estimated by the sample average of the subject-wise effects. When an effect is assumed homogeneous, the mean effect estimate records the marginal effect estimate.
3. A superscript † over the R^2 value indicates that the restricted default model (R_r^2) is rejected in favor of the unrestricted model (R_u^2) at the conventional significance levels by the χ_q^2 test of $(N-\kappa)(R_u^2 - R_r^2)/(1-R_u^2)$, where N is the sample size, κ the number of parameters in the unrestricted model, and q the difference in the numbers of parameters in the restricted and unrestricted models.