SMU ECONOMICS & STATISTICS
WORKING PAPER SERIES

SMU
SINGAPORE MANAGEMENT
UNIVERSITY

The WTO Trade Effect

Pao-li Chang, Myoung-jae Lee
September 2007

Paper No. 06-2007

The WTO Trade Effect

Pao-Li Chang*
School of Economics
Singapore Management University
Singapore 178903

Myoung-Jae Lee†
Department of Economics
Korea University
Seoul, Korea

September 23, 2007

Abstract

Rose (2004) showed that the WTO or its predecessor, the GATT, did not promote trade, based on conventional econometric analysis of gravity-type equations of trade. We argue that conclusions regarding the GATT/WTO trade effect based on gravity-type equations are arbitrary and subject to parametric misspecifications. We propose using nonparametric matching methods to estimate the 'treatment effect' of GATT/WTO membership, and permutation-based inferential procedures for assessing statistical significance of the estimated effects. A sensitivity analysis following Rosenbaum (2002) is then used to evaluate the sensitivity of our estimation results to potential selection biases. Contrary to Rose (2004), we find the effect of GATT/WTO membership economically and statistically significant, and far greater than that of the Generalized System of Preferences (GSP).

Keywords: GATT/WTO; GSP; treatment effect; matching; permutation test; signed-rank test; sensitivity analysis

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

*Tel: +65-6828-0830; fax: +65-6828-0833. *E-mail address:* plchang@smu.edu.sg (P.-L. Chang).
†Tel: +85-2-3163-4300; *E-mail address:* myoungjae@korea.ac.kr

1. INTRODUCTION

In the post war era, world merchandize trade volume grew at an exponential rate and at a rate much higher than world merchandize output: while exports volume grew at an annual rate of 6.2% during 1950-2005, production grew at only 3.8% (WTO, 2006, p. 27). This trend of expanded trade concurred with the development of the multilateral trade institution. Since the coming into effect of the General Agreement on Tariffs and Trade (GATT) in 1947, the framework of trade agreements regulating trade policies has grown in coverage and in depth over the decades, culminating with the establishment of the World Trade Organization (WTO) in 1995. The number of countries choosing to join the GATT/WTO also saw a steady increase from the original 23 contracting parties to the current 151 members. By contrasting the post-war phenomenon with the pre-war experience in the late 1920's to 1930's during which the world trade volume fell into a downward spiral with uncontrolled tariff retaliations among countries, it is natural for one to conclude that the GATT/WTO likely has spurred the integration and expansion of world trade.

On theoretical grounds, several economic theories have also been proposed that explain the mechanism of multilateral trade agreements in promoting trade. Major explanations include the terms-of-trade argument and the political-commitment argument. On one hand, multilateral trade agreements may help coordinate countries' trade policies and prevent them from engaging in the terms-of-trade-driven tariff retaliations which decrease trade volumes (Johnson, 1953–1954; Bagwell and Staiger, 1999, 2001). On the other hand, multilateral trade agreements may also help national governments to commit to liberalizing trade policies, given the retaliation threat from other countries if the commitment is not carried out. This enhances policy credibility with respect to domestic private sectors and brings about efficient production and trade structure (Staiger and Tabellini, 1987, 1989, 1999). Other potential mechanisms of GATT/WTO in encouraging trade often cited by economists or policymakers also include the reduced uncertainty of member countries' trade policies and overall business environment, facilitated by binding tariffs or the GATT/WTO dispute settlement mechanism (Evenett and Braga, 2005).

In view of the empirical observations and theoretical validation, it is thus surprising that some recent studies showed that the GATT/WTO did not appear to have promoted trade. Perhaps the most important and widely cited study is Rose (2004). The study uses a large

panel of data on bilateral trade for 178 trading entities during year 1948 to 1999, and conducts conventional econometric analysis based on the gravity model of trade (which postulates that bilateral trade depends log-linearly on the distance between two countries and their GDP's, and other determinants possibly affecting bilateral trade). The mechanism of GATT/WTO in influencing trade is incorporated in the framework by two dummy variables, which indicate respectively whether or not both countries are GATT/WTO members and whether or not only one country is a GATT/WTO member. The coefficients on both dummy variables were found in Rose (2004) to be insignificantly different from zeros, in the benchmark model with ordinary-least-squares (OLS) regression, as well as in most subsequent sensitivity analysis with variations in estimation methods and in the data.

Gravity-type models of trade have in recent years become popular in empirical studies of bilateral trade pattern. Although the basic formulation can be justified by theories as in Anderson (1979), Helpman and Krugman (1985), Deardorff (1998), and Anderson and van Wincoop (2003), the augmented version of gravity equation adopted in Rose (2004) (and in many other empirical studies) is often not. Various *ad hoc* regressors in addition to the basic gravity variables (the distance and the GDP's) are often added to the equation to suit the purpose of the study without rigorous theoretical justifications. For example, the survey by Oguledo and MacPhee (1994) cited 49 explanatory variables that had been used in gravity-type empirical studies. Although the extra variables typically used to measure the degree of trade frictions across countries (such as language barrier and colonial relationship) might well have an impact on bilateral trade flows, it is not clear in what functional form they would enter the basic gravity equation and how they would interact with one another. To illustrate this point, we show in the main text that the main gravity equation of Rose (2004) is misspecified by omitting higher-order interaction terms among regressors. Given this, it is possible to construct richer parametric forms of specifications and find favorable evidences of GATT/WTO's trade-creating effect. We provide one such example in the main text. However, without theoretical guidance, an investigation of this kind looks endless and any conclusion drawn would seem arbitrary.

In this paper, we propose using nonparametric 'matching' methods to estimate the trade effect of GATT/WTO membership. Nonparametric estimation does not assume knowledge of the underlying true model and avoids potential biases stemming from parametric misspecifications. If the gravity equation specified in Rose (2004) is indeed the correct underlying

model, using a nonparametric approach should lead to similar estimation results. Among many possible nonparametric approaches to evaluating the 'treatment' effect of a policy or program, matching method is relatively straightforward to implement. Matching is widely used in the fields of labor and health economics. See, for example, Heckman et al. (1997), Imbens (2004), and Lee (2005), and applications in Heckman et al. (1998), Lechner (2000), Lee et al. (2007), and Lu et al. (2001).

Applying the terminology in the matching literature to our current context, the state of both countries being GATT/WTO members is a 'treatment', and the country-couple in the state is a 'treated' subject, in contrast with a 'control' subject where the two countries are nonmembers. In essence, a pair-matching method compares the 'responses' (trade flows) of a treated subject and a control subject who differ in their treatments but are otherwise similar in every aspect likely to influence the bilateral trade flows. The difference in their bilateral trade flows is then attributed to the treatment effect of both countries being GATT/WTO members. Similar procedures apply if we want to estimate the effect of other treatments such as only one country in a country-couple being a GATT/WTO member versus both of them nonmembers. In finding the best match from the control group for a treated subject, a list of covariates need to be identified in terms of which the similarity between any two subjects is measured. For this, we use the same set of covariates as in Rose (2004). By using the matching method, however, we do not assume any particular functional form that relates these covariates to bilateral trade.

To assess the statistical significance of the treatment effect estimated by the matching method, we propose using the *permutation test*. Although matching estimators are popular in practice, asymptotic theories on their large sample properties are not fully established for most data generating processes. See Abadie and Imbens (2006) for a large sample theory of the matching estimator for independent observations. In practice, a standard t-statistic or a bootstrap procedure is often used to derive the p-value or confidence intervals. Standard t-statistic is straightforward but without theoretical justifications; bootstrap is computationally demanding and argued by Abadie and Imbens (2006) to be invalid. In contrast with the above asymptotics-based tests, the permutation test is an exact inferential procedure conditional on the observed data. Under the null hypothesis of no treatment effect, two subjects in a matched pair are 'exchangeable' in the labeling of their treatment status (treated or untreated). By obtaining all possible permutations of the treatment labels in all pairs, the exact p-value of

the observed matching estimator can be computed by placing it in the distribution of the permutations. See Pesarin (2001) and Good (2004) for an introduction to the concept, and Ernst (2004) for a survey of the permutation method in general contexts. Recently, Imbens and Rosenbaum (2005) showed that confidence intervals constructed by the permutation method are far more reliable than other confidence intervals.

In applying the matching estimator and the permutation test, there are two potential shortcomings. First, the matching estimator is susceptible to outliers (as OLS is). Second, the permutation test is only 'conditionally distribution-free' — conditional on the sample at hand; thus its finding is only valid within the sample. To provide more robust findings, we also conduct the Wilcoxon (1945) *signed-rank test* version of the matching estimator and the permutation test. Instead of the actual numerical differences, the ranks of the differences are used in determining the distribution of the signed-rank test statistic across all permutations. In addition to being robust to outliers, the signed-rank test turns out to be distribution-free, so that findings based on this test are also valid out of sample.

Some country-couples may be more likely to join the GATT/WTO and also to experience higher bilateral trade volumes, leading to a potential upward bias in the treatment effect estimator. This selection-bias problem is relevant in both the parametric setting of Rose (2004) and the nonparametric setting of ours. Selection on observable variables does not pose a selection-bias problem. If, however, the selection is made based on unobservables that affect both trade flows and the decision to join the GATT/WTO, then there is a selection problem. We conduct a sensitivity analysis following Rosenbaum (2002) to evaluate the sensitivity of our conclusions obtained under the assumption of no selection bias. This method examines how severe the unobserved selection problem must be to overturn the original finding. A finding is deemed robust if it takes a substantial difference between the treated and the untreated in their odds of taking treatment (after controlling for the observable variables) to overturn the finding. The use of similar sensitivity analysis is slowly increasing in numbers. See, for example, Aakvik (2001), Imbens (2003), Hujer et al. (2004), Altonji et al. (2005), and Lee et al. (2007), among others.

As another way to deal with the selection problem, we also apply the matching procedure introduced above in restricted manners, where the potential matches for a treated subject is limited to a certain subset of controls. They are: matching restricted to the observations of the same country-couple, matching restricted to the observations in the same year, matching

restricted to the observations in the same GATT/WTO period, and matching restricted to the observations of the same income-class combination. For example, there may be some unobserved country-couple specific characteristics that affect both their GATT/WTO memberships and their bilateral trade flows, and the characteristics may be systematically different across country-couples. In this case, matching without restriction leads to a biased estimate that contains the true treatment effect and the effect arising from the difference in unobserved country-couple characteristics if the pair in a match are of different country-couples. The bias is eliminated by restricting matching to the observations of the same country-couple.

Our nonparametric matching method and permutation-based inference procedures led to findings opposite to Rose's (2004): in short, we found economically and statistically significant trade-promoting effect of the multilateral trade institution. The bilateral trade between two countries is higher by 193% for country-couples who are both GATT/WTO members, relative to their counterparts who are both nonmembers. The effect is similar if estimation is restricted to cross-section variation (in the case of matching within the same year or the same GATT/WTO period). On the other hand, when estimation is restricted to time-series variation (in the case of matching within the same country-couple), the bilateral trade increases by a factor of 126% when a country-couple are both GATT/WTO members, relative to when they both are not. Both estimates are robust to a reasonable degree of potential selection biases. In his findings, Rose (2004) also suggested that the Generalized System of Preferences (GSP) had a significant and larger effect than GATT/WTO. In contrast, GSP is found in our analysis to raise bilateral trade volume, but by a factor smaller than the effect of 'both in GATT/WTO', regardless of the matching criteria. Based on matching restricted to country-couples of the same income-class combination, bilateral trade is estimated to be higher by a factor of 73% for a trading relationship that extends the GSP scheme, relative to its counterpart that does not.

The findings of Rose (2004) were also challenged by two other studies. On one hand, Tomz et al. (2005) highlight the effective participation of nonmember participants in the GATT/WTO system. These include colonies, *de facto* members, and provisional members, who enjoy to a large extent the same set of rights and obligations under the agreement as formal members. By simply re-classifying these countries as *de facto* GATT/WTO members without changing the estimation framework of Rose (2004), Tomz et al. (2005) find that GATT/WTO substantially increases trade and that its effects are relatively stable across

countries and over time. Subramanian and Wei (2007), on the other hand, emphasize asymmetries of GATT/WTO trade effects across different sub-samples—developed versus developing countries, old versus new developing countries, and generally unprotected sectors versus protected sectors. They find that the GATT/WTO institution promotes trade when and where the multilateral trade agreements are in full force. In contrast with these two papers, we do not attempt to refine the data of Rose (2004) or to differentiate the trade effects among sub-samples. Instead, our paper focuses on the improvement of estimation and inferential methodologies as introduced above.

The rest of this paper is organized as follows. In Section 2, we highlight possible misspecifications of the Rose (2004) gravity equation and arbitrariness of conclusions derived from the parametric framework regarding the GATT/WTO effect. We then introduce the nonparametric matching methodology and permutation-based inference procedures in Section 3. Section 4 reviews the data set of Rose (2004). Our estimation results and findings are reported in Section 5. Section 6 concludes.

2. MISSPECIFICATION OF THE GRAVITY EQUATION

In this section, we illustrate the potential misspecification of the main gravity equation used in Rose (2004) and the arbitrariness of conclusions derived based on variations of the model. The benchmark result of Rose (2004) based on an OLS regression of a gravity-type equation is replicated in Column (I) of Table 1.[1] As shown, the coefficients of the two GATT/WTO dummy variables (Both in and One in) are insignificantly different from zeros.

In Column (II), we run a misspecification test of the gravity equation (I). For this, we apply the regression equation specification error test (RESET) of Ramsey (1969), which tests whether non-linear combinations of the regressors have any power in explaining the regressand. If yes, then the original model is misspecified. The test is done by augmenting the original regression equation with the square or higher powers of the predicted value of the regressand from the original regression and testing the significance of the coefficients of these augmented terms. As indicated in Column (II), the coefficient of the squared predicted value $\hat{y}^2$ of the log real trade derived from the gravity equation (I) is significant. This indicates that the original gravity equation is misspecified by omitting interaction terms among regressors

[1] The data set used is Rose's (2004) downloaded from his website. More details on the data set are discussed in Section 4.

or squares of regressors, because $\hat{y}^2$ is a sum of these terms.

Given the finding of the RESET test, we experiment with augmenting the gravity equation (I) with interaction terms among the two GATT/WTO dummy variables and the other regressors, and rerun the regression. The results in Column (III) show that many interaction terms are significant, illustrating clearly that the benchmark model of Rose (2004) is misspecified.[2] With the interaction terms in, the effects of GATT/WTO membership for trading partners i and j in year t are,

$$\begin{aligned}
\{\text{Both in effect}\}_{ijt} &= \beta_{\{\text{Both in}\}} + \beta_{\{\text{Both in x GSP}\}}\{\text{GSP}\}_{ijt} + \dots \\
&+ \beta_{\{\text{Both in x Ever colony}\}}\{\text{Ever colony}\}_{ijt}; \\
\{\text{One in effect}\}_{ijt} &= \beta_{\{\text{One in}\}} + \beta_{\{\text{One in x GSP}\}}\{\text{GSP}\}_{ijt} + \dots \\
&+ \beta_{\{\text{One in x Common colonizer}\}}\{\text{Common colonizer}\}_{ijt} \\
&+ \beta_{\{\text{One in x Ever colony}\}}\{\text{Ever colony}\}_{ijt},
\end{aligned}$$

where β is the coefficient of the corresponding regressor or interaction term. These effects vary across trading partners and years ijt, and Table 2 shows some quantiles of these effects. As shown in Table 2, the mean effect of GATT/WTO membership for the whole sample based on the specification in (III), obtained by replacing the regressors with the sample means and β with the estimate, is significant and positive.

Of course, the specification in (III) is just one of many possibilities to allow higher-order parametric specifications of bilateral trade. One can conjecture that possibly richer parametric forms of specifications may lead to even more favorable evidence of GATT/WTO effect. However, without theoretical guidance, an investigation of this kind looks endless and any conclusion drawn would seem arbitrary.

In the following, we resort to the nonparametric matching approach to the estimation of the GATT/WTO trade effect. The approach requires identification of the covariates that are likely to have affected bilateral trade, for which we will adopt the same set of covariates as in Rose (2004). However, beyond that, no previous knowledge about the underlying model structure is required: estimation and inference are derived without having to specify the

[2]The null hypothesis that all interaction terms in parametric specification (III) are irrelevant can be easily rejected by the χ^2-test: $[(R_a^2 - R_0^2)/(1-R_a^2)] * (N-k)$, where R_a^2 (respectively R_0^2) is the R^2 under specification (III) (respectively (I)), N is the sample size, and k is the number of regressors under specification (III). The tiny difference in R^2 between parametric specifications (I) and (III) is magnified by $N-k$; as the current sample size is huge, we get significant interaction terms in spite of little change in the goodness-of-fit.

functional form that relates the bilateral trade pattern to the covariates; thus, all forms of higher-order or interaction terms of the covariates are accommodated. This avoids the often large variation in estimates across different functional forms and parametric modeling assumptions, and as a result, the biases stemming from model misspecifications.

3. METHODOLOGY

Suppose there are N countries in the world. With the countries indexed by $i = 1, ..., N$, let y_{ijt} be the logarithm of the real bilateral trade volume between countries i and j in year t. Define x_{ijt} as a vector of covariates, which includes mainly the characteristics of country couple (i, j) in year t. For country couple (i, j), two effects are of interest:

- the effect when both countries are in the GATT/WTO, relative to when none is in.
- the effect when only one country is in the GATT/WTO, relative to when none is in.

We will also investigate the effect of GSP on bilateral trade, which is:

- the effect when one country is a GSP beneficiary of the other country or vice versa, relative to the scenario of no GSP.

To facilitate exposition, focus on both-in versus none-in, and define

$$d_{ijt} = 1 \text{ if both in and } 0 \text{ if none in.}$$

The analysis for the effect of one-in or GSP is analogous. Thus, the data consists of $z_{ijt} = (d_{ijt}, x'_{ijt}, y_{ijt})'$, $i = 1, \ldots, N$, $j = i + 1, \ldots, N$, and $t = 1, \ldots, T$. In the terminology of the matching literature, y_{ijt} is a response variable, the observations with $d_{ijt} = 1$ constitute the treatment group, and those with $d_{ijt} = 0$ the control group. Note that we will use the word 'couple' (an observation unit which comprises two countries between which trade volume is concerned) and 'pair' (two observation units which are considered a match) separately to avoid confusion with the matching literature.

3.1 Mean Effects and Matching

Define two 'potential' response variables:

$$\begin{aligned} y^1_{ijt} &: \text{ potential } \textit{treated} \text{ response for couple } (i,j) \text{ in year } t, \\ y^0_{ijt} &: \text{ potential } \textit{untreated} \text{ response for couple } (i,j) \text{ in year } t. \end{aligned}$$

We will denote 'couple (i,j) in year t' simply 'subject ijt' from now on. The subject-ijt treatment effect is $y^1_{ijt} - y^0_{ijt}$, which is, however, not identified. Instead, usually one tries to estimate the mean effect $E(y^1_{ijt} - y^0_{ijt})$. Omitting the subscripts to simplify exposition, $E(y^1 - y^0)$ is identified by the group mean difference

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

where 'II' stands for statistical independence. If we have only $y^0 \amalg d$, then

$$\begin{aligned} & E(y|d=1) - E(y|d=0) = E(y^1|d=1) - E(y^0|d=0) = E(y^1|d=1) - E(y^0|d=1) \\ = \;& E(y^1 - y^0|d=1) \quad \text{which is the '}\textit{effect on the treated}\text{'.} \end{aligned}$$

Analogously, if $y^1 \amalg d$, then

$$E(y|d=1) - E(y|d=0) = E(y^1 - y^0|d=0), \quad \text{which is the '}\textit{effect on the untreated}\text{'.}$$

Since y^0 is the control response, in general, $y^0 \amalg d$ is more plausible than $y^1 \amalg d$.

For the GATT/WTO effect, both the effect on the treated (i.e., those who chose to join GATT/WTO) as well as the effect on the untreated (i.e., those who chose not to join) are of interest. Since

$$E(y^1 - y^0) = E(y^1 - y^0|d=0)P(d=0) + E(y^1 - y^0|d=1)P(d=1),$$

if the effect on the treated is the same as the effect on the untreated, then both equal $E(y^1 - y^0)$. Otherwise, the representative effect $E(y^1 - y^0)$ is a weighted average of the effect on the treated and the effect on the untreated.

In observational data, the treatment and control groups usually differ in observed (x_{ijt}) or

unobserved variables (ε_{ijt}). Matching removes the observed difference (and the unobserved difference to the extent the two differences are related). In terms of equations, matching is expressed by including x in the conditioning set in the preceding equations. The above group mean difference equation becomes

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

and the $E(y^1-y^0)$-decomposition equation becomes

$$E(y^1-y^0|x) = E(y^1-y^0|d=0,x)P(d=0|x) + E(y^1-y^0|d=1,x)P(d=1|x).$$

Once the x-conditional effect is found, x can be integrated out to yield a marginal effect; here, the integration can be done with different integrators (i.e., weighting functions). For the effect on the treated, the distribution of $x|d=1$ is typically used to render

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

where $F(\cdot|d=1)$ denotes the distribution of $x|d=1$.

The assumption $(y^0,y^1)\amalg d|x$ states that y^0 and y^1 may be related to d but only through x. This is nothing but a 'no-selection bias' assumption, which may be called 'randomization of d given x'. In the current context of bilateral trade, this assumption implies that a treated subject and an untreated subject be comparable in terms of their potential trade performance and likelihood of joining the GATT/WTO, if they exhibit the same observable characteristics x. Thus, the observed untreated response of the untreated subject can be used as a proxy for the counterfactual potential untreated response of the treated subject, and the observed treated response of the treated subject a proxy for the counterfactual potential treated response of the untreated subject.

If we are only concerned with the effect on the treated, $E(y^1-y^0|d=1)$, the required assumption, $y^0 \amalg d|x$, is weaker and requires only that a treated subject and an untreated subject be comparable in terms of their potential *untreated* trade performance and likelihood of joining the GATT/WTO, if they exhibit the same observable characteristics x. Again, this implies that the observed untreated response of the untreated subject can be used as a proxy for the counterfactual potential untreated response of the treated subject. The mean

effect on the treated can then be estimated based on the subset of the sample who receive the treatment. The opposite is true if we are only concerned with the effect on the untreated.

Thus, the assumption for matching to work in essence requires that there be no systematic unobserved difference across the treatment and control group that affects their potential (*treated* or *untreated* or both) trade volumes and GATT/WTO member status. If there is an unobserved variable ε that affects both d and (y^0, y^1), then there is a selection problem and ε causes a bias—called 'hidden bias'. Exactly for the unknown nature of the treatment effect and its unknown dependence on the unobserved characteristics, in general, it is more plausible for the assumption $y^0 \amalg d|x$ to hold than $y^1 \amalg d|x$. Thus, in the empirical section, we will place more emphasis on the estimation result of the effect on the treated than either the effect on the untreated or the mean effect for the whole sample. While any effect of x on (y^0, y^1) and d is dealt with by matching, there is no good way to deal with ε, for it is unobserved. Nevertheless, a sensitivity analysis can be conducted to account for the presence of ε and the extent it influences d and (y^0, y^1) as to be shown later.

So far, we introduced the basics of the treatment effect framework and matching concept. In practice, matching for the effect on the treated proceeds in the following steps (the effect on the untreated can be handled analogously). First, a treated unit, say unit ijt, is selected. Second, control units are selected who are the closest to the treated unit ijt in terms of x. If only one unit is selected, we get a 'pair-matching'; otherwise, multiple-matching, in which the number of matched controls may be allowed to be random or fixed. Third, assuming pair-matching, suppose there are M pairs (where M corresponds to the number of the successfully matched treated subjects), and y_{m1} and y_{m2} are the responses of the two subjects in pair m ordered such that $y_{m1} > y_{m2}$ without loss of generality; drop any pair with $y_{m1} = y_{m2}$. Then, defining

$$s_m = 1 \text{ if the first subject in pair } m \text{ is treated and } -1 \text{ otherwise,}$$

the effect on the treated can be estimated with a pair-matching estimator

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

which is simply the average of the difference between the treated response and the untreated

response across the M pairs of match.

Some remarks are in order. First, the matching scheme can be reversed to result in an estimator for the effect on the untreated: a control unit is selected first and then a matched unit from the treatment group later. Second, the distance $\|x_{ijt} - x_{i'j't'}\|$ between two units ijt and $i'j't'$ in terms of x should be chosen; e.g., a scale-normalized distance between x_{ijt} and $x_{i'j't'}$, or alternatively, $(x_{ijt} - x_{i'j't'})' X_N^{-1} (x_{ijt} - x_{i'j't'})$ can be used, where X_N is the sample variance matrix in the pooled sample. Third, for a treated unit, if there is no good matched control, then the unit may be passed over; i.e., a 'caliper' c may be set such that a treated unit with

$$\min_{i'j't' \in C} \|x_{ijt} - x_{i'j't'}\| > c, \quad \text{where '} i'j't' \in C\text{' means unit } i'j't' \text{ in the control group } C,$$

is discarded. The number of matched pairs M used in the matching estimator D decreases accordingly. For more discussions on treatment effect and matching in general, see, for example, Rosenbaum (2002) and Lee (2005).

3.2 Permutation Test for Matched Pairs

Suppose that matching has been done resulting in M pairs, and that the two subjects are *exchangeable* under the null hypothesis H_0 of no effect: the joint distribution of the two responses in each matched pair does not change when the two responses are switched. That is, exchangeability is taken as the definition of 'no effect', which is weaker than the subject-wise no-effect concept $y^1_{ijt} - y^0_{ijt} = 0$, but stronger than the mean zero-effect concept $E(y^1_{ijt} - y^0_{ijt}) = 0$. Exchangeability implies that, given the data, all 2^M possibilities to permute two responses in each pair are all equally likely with probability 2^{-M} in the 'permutation sample space' with the 2^M elements (equivalently, if they are not equally likely, then exchangeability does not hold). Here, permuting two responses is equivalent to assigning one response to the treatment group and the other to the control group.

Permutation inference is an exact inferential procedure conditional on all observed data. Whether H_0 is rejected or not depends on how extreme D is—i.e., the p-value of D. In theory, this can be computed exactly as

$$\frac{1}{2^M} \sum_{k=1}^{2^M} 1[D_k > D] \quad \text{if the } H_0\text{-rejection region is in the upper tail}$$

where D_k is a "D-like" effect estimator under permutation k. In practice, however, finding the p-value this way tends to be too time-consuming. Instead, we will apply the following sequence of approximations.

Generate M *iid* random variables w_m, $m = 1, ..., M$, with $P(w_m = 1) = P(w_m = -1) = 0.5$. If $w_m = -1$, the two responses in pair m are switched (the treated response assigned to the control group, and the untreated response to the treatment group); otherwise, the two responses in pair m are not switched. For pseudo sample k obtained this way, obtain the pseudo effect estimator

$$D'_k \equiv \frac{1}{M}\sum_{m=1}^{M} w_m s_m (y_{m1} - y_{m2})$$

where only w_m's are random. The p-value of D is then

$$\frac{1}{K}\sum_{k=1}^{K} 1[D'_k \quad > \quad D] \quad \text{if the } H_0\text{-rejection region is in the upper tail,}$$

where K is a number much smaller than 2^M, say $K = 1000$. In principle, a pseudo sample should not be repeated: 'sampling from the 2^M possibilities without replacement' is required. In practice, the chance of a pseudo sample being repeated in K $(= 1000)$ pseudo samples is negligible and can be ignored.[3]

A further simplification is possible by applying the central limit theorem (CLT) to w_m's. Observe

$$E(D'_k) = 0 \quad \text{and} \quad V(D'_k) = E(D'^2_k) = \frac{1}{M^2}\sum_{m=1}^{M} E\{w_m^2 s_m^2 (y_{m1} - y_{m2})^2\} = \frac{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2}{M^2}.$$

When M is large, the normally approximated p-value of D is

$$\begin{aligned} P(D'_k \quad > \quad D) &= P\{\frac{D'_k}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}} > \frac{D}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}}\} \\ &= P\{N(0,1) > \frac{D}{\{\sum_{m=1}^{M}(y_{m1} - y_{m2})^2/M^2\}^{1/2}}\}, \end{aligned}$$

if the H_0-rejection region is in the upper tail. Using this normal approximation, one does not even have to simulate w_m's.

As is well known, we can obtain a confidence interval (CI) by "inverting" the test pro-

[3]The proof is available from the authors upon request.

cedure. For instance, suppose that the treatment effect is the same for all pairs: the effect increases the treated response by a constant, say β. In this case, replace y_{m1} with $y_{m1} - \beta$ when $s_m = 1$ or y_{m2} with $y_{m2} - \beta$ when $s_m = -1$ to obtain

$$D_\beta \equiv \frac{1}{M} \sum_{m=1}^{M} s_m(y_{m1} - s_m\beta - y_{m2})$$

and restore the no-effect situation. Define accordingly

$$D'_\beta \equiv \frac{1}{M} \sum_{m=1}^{M} w_m s_m(y_{m1} - s_m\beta - y_{m2})$$

to observe

$$E(D'_\beta) = 0 \text{ and } V(D'_\beta) = E(D'^2_\beta) = \frac{\sum_{m=1}^{M}(y_{m1} - s_m\beta - y_{m2})^2}{M^2}.$$

Now conduct level-α tests with

$$\frac{D_\beta}{\{\sum_{m=1}^{M}(y_{m1} - s_m\beta - y_{m2})^2/M^2\}^{1/2}}$$

as β varies. The collection of β values that are not rejected is the $(1-\alpha)100\%$ confidence interval for β.

See Martiz (1995), Hollander and Wolfe (1999), and Lehmann and Romano (2005) among many others, for more on permutation (or randomization) tests in general. The use of permutation-based tests, in stead of asymptotic tests, is especially convenient in the current context with a panel of bilateral trade data, which possibly have a complicated data structure with serial and spatial dependence, rendering the derivation of the asymptotic distribution of the matching estimator difficult. Based on the principle of exchangeability, the permutation test requires the joint distribution of the treated and untreated responses to remain the same despite the permutation. If the joint distribution is normal, this holds if the x-conditional variances of the two responses are the same under the null of no effect. This allows for any form of correlation between a treated response and an untreated response in any matched pair and across matched pairs, and any form of heteroskedasticity across matched pairs, thus accommodating a wide range of data structure.

3.3 Signed-Rank Test for Matched Pairs

The preceding permutation test has two potential shortcomings. One is its susceptibility to outliers in x and y. The other is that the test is only 'conditionally distribution-free'—the permutation distribution under the null hypothesis is conditional on (x', y, d), which enters the variance $V(D'_k)$. Due to this conditioning, any finding from the test is applicable only to the sample at hand. That is, the finding has 'internal validity', but not 'external validity'.

If a test is distribution-free, then the finding holds unconditionally as well, which accords external validity. A well known test that is robust and distribution-free is the Wilcoxon (1945) *signed-rank test*. Applying this test to the current matching context, rank $|y_{m1} - y_{m2}|$, $m = 1, ..., M$, to denote the resulting ranks as $r_1, ..., r_M$. For instance, when $M = 3$,

$$|y_{11} - y_{12}| = 0.3,\ |y_{21} - y_{22}| = 0.09,\ |y_{31} - y_{32}| = 0.21 \implies r_1 = 3,\ r_2 = 1,\ r_3 = 2.$$

The signed-rank test statistic is the sum of the ranks for the pairs where the treated subject has the higher response:

$$R \equiv \sum_{m=1}^{M} r_m 1[s_m(y_{m1} - y_{m2}) > 0] = \sum_{m=1}^{M} r_m 1[s_m = 1].$$

Thus, instead of actual numerical differences, the ranks of the differences are used in constructing the signed-rank R statistic, rendering it more robust to outliers than the estimator D.

The inference based on the signed-rank R statistic can similarly follow the permutation inferential procedure. The permutated version R' for R is

$$\begin{aligned} R' &\equiv \sum_{m=1}^{M} r_m 1[w_m s_m(y_{m1} - y_{m2}) > 0] = \sum_{m=1}^{M} r_m 1[w_m s_m > 0] \\ &= \sum_{m=1}^{M} r_m (1[w_m = 1, s_m = 1] + 1[w_m = -1, s_m = -1]). \end{aligned}$$

Observe

$$\begin{aligned} E(1[w_m &= 1, s_m = 1] + 1[w_m = -1, s_m = -1]) = \frac{1[s_m = 1]}{2} + \frac{1[s_m = -1]}{2} = \frac{1}{2}. \\ V(1[w_m &= 1, s_m = 1] + 1[w_m = -1, s_m = -1]) = \frac{1}{4}. \end{aligned}$$

Hence, because r_m's are fixed, under the H_0,

$$\begin{aligned} E(R') &= \sum_{m=1}^{M} r_m \frac{1}{2} = \frac{1}{2}\sum_{m=1}^{M} r_m = \frac{M(M+1)}{4}, \\ V(R') &= \sum_{m=1}^{M} r_m^2 \frac{1}{4} = \frac{1}{4}\sum_{m=1}^{M} r_m^2 = \frac{M(M+1)(2M+1)}{24}. \end{aligned}$$

When M is large, the null distribution of $\{R' - E(R')\}/SD(R')$ can be approximated by $N(0,1)$. The normally approximated p-value for R is

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

This test is distribution-free as (x', y, d) does not appear in the mean or variance of the null distribution.

The CI's for the treatment effect can be similarly obtained by inverting the test. Conduct level-α tests with different values of β using

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

and $r_{m\beta}$ is the rank of $|y_{m1} - s_m\beta - y_{m2}|$, $m = 1, ..., M$.

In contrast to the confidence interval obtained based on D_β with its point effect estimator D, there is no point effect estimator available in the signed-rank test. As a point estimator, one may take the Hodges and Lehmann (1963) estimator, which is obtained by solving for β such that

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

That is, after transforming the data with the effect estimate β, one obtains a transformed signed-rank statistic R_β that coincides with the mean of the statistic under the null.

3.4 Sensitivity Analysis with Signed-Rank Test

So far the unobserved differences between two subjects in a matched pair have been ignored. Rosenbaum (2002) presents a sensitivity analysis to account for an unobserved confounder ε that might affect d. Since our data set is observational where countries self-select the treatment, there may be unobserved differences across the treatment and control groups,

causing a hidden bias (or selection bias). Besides, not all country-couples are observed at all years; this is a missing variable problem, which is also a selection problem. Hence it matters to allow for selection problems.

When two subjects in a given pair differ in ε that influences d, their relative probabilities of receiving the treatment are no longer the same. Rosenbaum (2002) assumes

$$\frac{1}{\Gamma} \leq \frac{\text{odds of one subject being treated}}{\text{odds of the other being treated}} \leq \Gamma, \quad \text{for some constant } \Gamma \geq 1 \ \forall \text{ pair.}$$

For instance, if the first subject's probability of receiving the treatment is 0.6 (0.6) and the second subject's probability is 0.5 (0.4), then the odds ratio is

$$\frac{0.6/0.4}{0.5/0.5} = 1.5 \quad (\frac{0.6/0.4}{0.4/0.6} = 2.25).$$

The sensitivity analysis goes as follows. Initially, one proceeds under the assumption of no unobserved difference ($\Gamma = 1$). Then Γ is increased from 1 to see how the initial conclusion is affected. If it takes a large value of Γ to reverse the initial finding, i.e., if only a strong presence of ε can overturn the initial conclusion, then the initial conclusion is deemed insensitive to ε. Otherwise, if it takes only a small value of Γ, then the initial finding is deemed sensitive.

The question is then how "large" is large for Γ. Suppose $\Gamma = 2.25$ and imagine that ε were observed. By observing ε, one would be able to tell who is more likely to be treated between the two subjects and ask whether the probability difference results in as much as an odds ratio of 2.25? If the answer is no, $\Gamma = 2.25$ is a large value. Thus how large is large for Γ depends on what is included in x. If most relevant variables are in x and thus if it is hard to think of any important omitted variable in ε, then even a small value of Γ may be regarded as large.

An easy-to-implement sensitivity analysis is available for the signed-rank test. Define

$$p^+ \equiv \frac{\Gamma}{1+\Gamma} \geq 0.5 \quad \text{and} \quad p^- \equiv \frac{1}{1+\Gamma} \leq 0.5.$$

Further define R^+ (R^-) as the sum of M-many independent random variables where the mth variable takes 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

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

Doing analogously, we obtain

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

The means and variances with p^+ and p^- include $E(R')$ and $V(R')$ as a special case when $p^+ = p^- = 1/2$.

Suppose that the H_0-rejection interval is in the upper tail. Rosenbaum (2002, p.111) shows that

$$P(R^+ \geq a) \geq P(R' \geq a) \geq P(R^- \geq a).$$

Using these bounds, the p-value obtained under no hidden bias can be bounded by $P(R^+ \geq a)$ in case of rejection.[4] Specifically, suppose that the H_0-rejection interval is in the upper tail, and the no-hidden-bias p-value is

$$P\{N(0,1) \geq \frac{R - E(R')}{SD(R')}\} = 0.001,$$

leading to the rejection of H_0 at level $\alpha > 0.001$. Rewrite $P(R^+ \geq a) \geq P(R' \geq a)$ as

$$\begin{aligned} & P\{\frac{R^+ - E(R^+)}{SD(R^+)} \geq \frac{a - E(R^+)}{SD(R^+)}\} \geq P\{\frac{R' - E(R')}{SD(R')} \geq \frac{a - E(R')}{SD(R')}\} \\ \simeq\ & P\{N(0,1) \geq \frac{a - E(R^+)}{SD(R^+)}\} \geq P\{N(0,1) \geq \frac{a - E(R')}{SD(R')}\}. \end{aligned}$$

Replacing a in the above equation with the realized R gives the p-value of R on the right hand side and its bound on the left hand side. The left-hand side can be obtained for different

[4] In case of acceptance, $P(R^- \geq a)$ shows the possibility of rejection when ε is taken into account. For a two-sided test, the p-value gets multiplied by 2. For a lower-tail test, subtracting the last display from 1 to get

$$1 - P(R^+ \geq a) \leq 1 - P(R' \geq a) \leq 1 - P(R^- \geq a) \implies P(R^+ < a) \leq P(R' < a) \leq P(R^- < a),$$

which can be used for bounds.

values of Γ. Then find, at which values of Γ, the upper bound crosses the level α. If this happens at, say $\Gamma = 2$, then check whether or not $\Gamma = 2$ is a large value. If $\Gamma = 2$ is deemed large, then the initial rejection is insensitive to the unobserved difference.

The relevant distribution (R^+ or R^-) to use for calculating the bound in the sensitivity analysis indicates the possible direction of selection bias that could undermine an initial significant finding of treatment effect. If the finding is a significantly positive effect, then we only need to worry about the 'positive' selection problem, 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 a reverse 'negative' selection problem, where a subject with a lower potential treatment effect is also more likely to be treated, can weaken the original finding, so the sensitivity analysis in the direction of R^- is applicable.

4. DATA

Since our focus is on improving the estimation methodology in Rose (2004), we use the original data set of Rose (2004) without further modifications or amendments.[5] Rose (2004) provides a detailed account of the construction of the data set.

The response variable y_{ijt} in our matching framework corresponds to the regressand in the Rose's (2004) gravity equation, which measures (the natural logarithm of) the average value of real bilateral trade between countries i and j in year t. The treatment dummy variable d_{ijt} corresponds to one of the three binary variables ($Bothin_{ijt}$, $Onein_{ijt}$, GSP_{ijt}). They measure, respectively, whether both countries i and j are GATT/WTO members in year t, whether only one of the two countries (i, j) is a GATT/WTO member in year t, and whether country i is a GSP beneficiary of country j or vice versa in year t.

The conditioning variables or covariates x_{ijt} in our matching framework correspond to the complete list of regressors in Rose's (2004) benchmark gravity equation. This includes: (1) (the natural logarithm of) the distance between countries i and j; (2) (the natural logarithm of) the product of real GDP's of the country couple (i, j) in year t; (3) (the natural logarithm of) the product of real per capita GDP's of the country couple (i, j) in year t; (4) a binary variable which indicates whether the country couple (i, j) share a common language; (5) a

[5]The data set is available from Rose's Web site (http://faculty.haas.berkeley.edu/arose/GATTdataStata.zip).

binary variable which indicates whether the country couple (i, j) share a land border; (6) a discrete variable which counts the number of landlocked countries in the country couple (i, j); (7) a discrete variable which counts the number of island nations in the country couple (i, j); (8) (the natural logarithm of) the product of land areas of the country couple (i, j); (9) a binary variable which indicates whether the country couple (i, j) were ever colonies after 1945 with the same colonizer; (10) a binary variable which indicates whether country i is a colony of country j in year t or vice versa; (11) a binary variable which indicates whether country i ever colonized country j or vice versa; (12) a binary variable which indicates whether the country couple (i, j) remained part of the same nation during the sample; (13) a binary variable which indicates whether the country couple (i, j) use the same currency in year t; (14) a binary variable which indicates whether the country couple (i, j) belong to the same regional trade agreement, (15) a list of year dummies for $t = 1948, \ldots, 1999$; and (16) two of the three binary variables ($Bothin$, $Onein$, GSP), with the one whose treatment effect is being investigated suitably excluded from the list.

The complete sample contains the above variables observed for 178 IMF trading entities between 1948 and 1999 (with gaps). This amounts to a total of 234,597 observations.

5. ESTIMATION AND RESULTS

5.1 Matching Criteria and Procedures

We begin with the baseline methodology described in Section 3. This approach, by matching in terms of the conditioning variables x, controls for the difference in bilateral trade arising from differences in observable characteristics across country-couples and years (such as geographical proximity and output levels). Although the set of conditioning variables x used for matching is quite extensive, it is possible that some systematic differences in unobservable characteristics ε remain which lead to selection bias. The sensitivity analysis introduced in Section 3.4 helps assess the sensitivity of estimation results to whatever selection bias may remain; it, however, does not address the source of selection bias. We experiment with different matching criteria to address various potential selection biases. They are: matching restricted to the observations of the same country-couple, matching restricted to the observations in the same year, matching restricted to the observations in the same GATT/WTO period, and matching restricted to the observations of the same income-class combination

(details of the above criteria to be elaborated later). By restricting the potential match to the observations of the specified criterion, we are accounting for the possibility that systematic unobservable differences exist (across country-couples, across years, across GATT/WTO periods, or across income-class combinations) which influence bilateral trade patterns and decisions to join the GATT/WTO or decisions to extend GSP. Tables 3–7 report the results for each set of matching criterion.

For each set of matching criterion, we estimate the treatment effect of: 'Both in GATT/WTO', 'One in GATT/WTO', and 'GSP', respectively. For the GATT/WTO effect, both the effect on the treated (i.e., those who chose to join the GATT/WTO) as well as the effect on the untreated (i.e., those who chose not to join) are of interest. We report both effects and the weighted average of them as the representative effect for all. For the GSP effect, we report only the effect on the treated. We do not report the effect on the untreated, as GSP's are unilateral trade preferences extended only from the rich to the developing country. It does not make much sense to propose a GSP arrangement between two rich countries, for example, and to investigate the potential trade effect. On the other hand, the GSP effect on the treated that we will report below should also be taken with a grain of salt, except in the last matching exercise where matching is restricted to the observations of the same income-class combination. This is because the decision to extend GSP's and the bilateral trade pattern may be both dependent on the relative income level of trading partners, which poses potential selection bias problems. The last matching exercise, with matching restricted to the observations of the same income-class combination, is least subject to this caveat.

When the treatment effect of 'Both in GATT/WTO' is investigated, observations with only one country in the GATT/WTO are dropped. This leaves a remaining sample with only observations where the two countries are either both in the GATT/WTO (114,750 observations) or both are outside the GATT/WTO (21,037 observations). Similarly, when the treatment effect of 'One in GATT/WTO' is investigated, observations with both countries in the GATT/WTO are dropped. The remaining sample thus includes only observations where either only one of the two countries is in the GATT/WTO (98,810) or both are outside the GATT/WTO (21,037). For the treatment effect of 'GSP', a small proportion (54,285) of the whole sample (234,597) has GSP arrangements.

In all matching exercises, we restrict our attention to the case of pair-matching, in which only one matched unit is selected. Since our conditioning variables x contain continuous

variables, the likelihood of two matched units having equal distance in terms of x from the target is negligible and the case of multiple-matching can be safely ignored. In order to measure the distance $\|x_{ijt} - x_{i'j't'}\|$, we use the simple scale-normalized distance.

In Section 3.2, we introduced both simulation and normal approximation approaches to obtaining the p-value of the treatment effect estimator D based on the permutation test. Although not mentioned explicitly in Section 3.3, in addition to the normal approximation approach, it is also possible to calculate the p-value for the signed-rank R statistic based on simulated permutation samples. We carried out both approaches and found the normal p-value to approximate the simulated p-value extremely well (which is expected as the sample size is large). Thus, we present only the normal approximation result in the reports that follow.

In any given matching exercise, we experiment with three caliper choices: the caliper is set such that only 100%, 80%, or 60% of matched pairs are qualified for the estimation of the treatment effect. For example, for the caliper choice of 60%, matched pairs with a distance (in terms of x) exceeding the upper 60 percentile of all matched pairs are discarded. The caliper choice of 100% is equivalent to using all available matched pairs.

5.2 Matching Results

In Tables 3–7, the first column 'permutation test' reports the results on permutation tests. The 'effect' sub-column presents the treatment effect estimate based on the D statistic, the p-value is obtained for the observed D statistic using the permutation test based on the normal approximation approach, and the CI is obtained by inverting the test procedure. The second column 'signed-rank test' reports the results on signed-rank tests. The 'effect' sub-column presents the treatment effect estimate based on the Hodges and Lehmann (1963) estimator, the p-value is obtained for the observed R statistic using the signed-rank test, and the CI is obtained by inverting the test procedure. The third column 'sensitivity analysis' reports the results on sensitivity analysis. The sensitivity analysis is conducted for the signed-rank test based on a significance level of $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of Γ) indicates the relevant distribution on which the bound for the p-value of the signed-rank test is based. Γ^* indicates the critical value of Γ at which the conclusion of the signed-rank test reverses.

We discuss Tables 3–7 now. Table 3 reports the results of the baseline methodology,

labeled 'unrestricted matching'. In this case, the pool of potential matches for an observation are the observations with the opposite treatment; no further restriction is imposed. The numbers of matched pairs obtained (M_1 for the case of the effect on the treated, and M_0 for the untreated) are specified in the parentheses. We note that the permutation test and the signed-rank test produce similar results. As the signed-rank test is considered more robust to outliers and is also directly related to the sensitivity analysis, we cite the signed-rank test's figures below for illustrative purpose.

First, we note that the treatment effect of 'Both in GATT/WTO' on the treated is large and significant. For example, with the 80% caliper, the point estimate indicates that the GATT/WTO membership raises bilateral trade volume by 193% ($= e^{1.075} - 1$) if both trading partners are GATT/WTO members. The corresponding effect is smaller at 76% ($= e^{0.568} - 1$) if only one joins the GATT/WTO. Nevertheless, the effect is still positive and significant. Thus, our estimation results suggest that the trade-creating effect of GATT/WTO membership does not come at the cost of diverting trade from nonmembers. Relative to the GATT/WTO membership, the preferential GSP scheme also promotes bilateral trade, by a factor of 101% ($= e^{0.696} - 1$) with the 80% caliper, for trading relationships that apply the scheme. It is important to note that the positive and stronger trade effect of 'Both in GATT/WTO' is shared by a larger number of bilateral trading relationships ($114,750$), than the GSP scheme ($54,285$). Thus, either on the average or in the aggregate, our estimation results suggest that the *realized* trade-creating effect of GATT/WTO membership exceeds that of GSP by a great extent.

How about the *potential* trade effect of GATT/WTO for trading relationships that are outside the system? The treatment effect estimates on the untreated (with the 80% caliper) suggest that the bilateral trade volume would have increased by 22% ($= e^{0.200} - 1$) if both trading partners were to join the GATT/WTO or 9% ($= e^{0.089} - 1$) if only one of them did so. The effects are smaller compared to the effect on the treated but they are still significantly positive. As discussed in Section 3.1, the required assumption for applying the matching estimator to the treated, $y^0 \amalg d|x$, is more likely to hold than to the untreated, $y^1 \amalg d|x$. Thus, we will place less emphasis on the effect on the untreated in the following discussions, although their results will continue to be reported in the tables.

Table 4 reports the results for 'matching within country-couple'. In this case, the pool of potential matches for an observation are restricted to the observations with the oppo-

site treatment and of the same country-couple. Some country-couples may be both in the GATT/WTO, be both outside the GATT/WTO, or have only one of them in the GATT/WTO, throughout the sampling years (1948 to 1999). Alternatively, some country-couples may have only one of them in the GATT/WTO for some period of the sampling years and then be both in the GATT/WTO throughout the rest of the sampling years. In these cases, these observations do not have qualified matched controls (or matched treatment), and are discarded from the sample. This explains the much smaller sample size shown in Table 4. The estimated treatment effects on the treated for the three types of treatments have the same ranking as in the case of unrestricted matching: *Bothin* effect > *GSP* effect > *Onein* effect (with the 80% caliper). While the current estimates suggest overall smaller treatment effects on the treated, the trade-creating effect if both trading partners are in the GATT/WTO continues to be economically (and statistically) significant: for example, based on the 80% caliper, bilateral trade increases by as much as 126% ($= e^{0.814} - 1$).

In Table 5, labeled 'matching within year', the pool of potential matches for an observation are restricted to the observations with the opposite treatment and in the same year. In this case, the year dummies in x are redundant and so are dropped from x. The estimation results differ slightly from those of the 'unrestricted matching' in terms of the permutation test, but they come across as almost the same as those in Table 3 in terms of the signed-rank test. This indicates that in the case of 'unrestricted matching', the matched pairs often occur among cross-section observations of the same year, and thus the estimates pick up mostly the cross-section variations, whereas the estimates derived from 'matching within country-couple' as discussed in the previous paragraph measure only time-series variations. Both cross-section and time-series variations indicate that there are significant gains in trade by joining the GATT/WTO.

Table 6 presents the results for 'matching within period'. The pool of potential matches for an observation are restricted to the observations with the opposite treatment and in the same period, where the periods correspond to different eras of the GATT/WTO history. They are: 1948 (Before Annecy round), 1949-1951 (Annecy to Torquay round), 1952-1956 (Torquay to Geneva round), 1957-1961 (Geneva to Dillon round), 1962-1967 (Dillon to Kennedy round), 1968-1979 (Kennedy to Tokyo round), 1980-1994 (Tokyo to Uruguay round), 1995-(After Uruguay round). Given our earlier observations that in the case of 'unrestricted matching', the matched pairs often occur among cross-section observations of the same year, it is no

surprise to see that the current estimation results again appear to be almost the same as in the case of 'unrestricted matching', as the criterion of matching within the same period in effect does not impose much extra restriction.

Table 7 reports the final set of results, for 'matching within income-class combination'. In this matching exercise, the pool of potential matches for an observation are restricted to the observations with the opposite treatment and of the same income-class combination. The income-class combinations are: 'low income-low income' country-couples, 'low income-middle income' country couples, 'low income-high income' country couples, 'middle income-middle income' country couples, 'middle income-high income' country couples, and 'high income-high income' country couples. Observations without a qualified match are discarded. The estimated treatment effects on the treated are overall smaller for GATT/WTO membership as well as for GSP, although their relative ranking remains the same: *Bothin* effect > *GSP* effect > *Onein* effect (with the 80% caliper), as with other matching criteria. The current estimates suggest that the GATT/WTO membership promotes bilateral trade, by an economically and statistically significant factor of: 108% ($= e^{0.734} - 1$) if both trading partners are in the GATT/WTO and 58% ($= e^{0.460} - 1$) if only one of them is in the GATT/WTO. As discussed earlier, the estimated GSP treatment effect with matching within the same income-class combination is likely to be less subject to the selection bias problem. Based on this setup, the GSP treatment effect is estimated to raise bilateral trade by a factor of 73% ($= e^{0.551} - 1$).

5.3 Sensitivity Analysis Results

How robust are the results we have referred to so far? By the sensitivity analysis, Table 3 indicates that in the case of 'unrestricted matching', the significant treatment effects of 'Both in GATT/WTO' and 'GSP' are robust to any positive selection bias that leads to as much as an odds ratio of 2.081, and 2.117 respectively, in receiving treatment between the treated and untreated subject in any matched pair (by the 80% caliper and the two-sided test). On the other hand, the robustness of the significant treatment effect of 'One in GATT/WTO' to potential positive selection bias is relatively weaker (measured by Γ^* at 1.521). The same degrees of robustness of the three treatments reappear in Tables 5 and 6, when the matching is restricted to the observations in the same year, or in the same period.

Are these figures large enough? In using similar sensitivity analysis, Aakvik (2001) seems to regard $\Gamma = 1.5 \sim 2$ as large, while Hujer et al. (2004) appear to base their discussions

on lower numbers of $\Gamma = 1.25 \sim 1.5$. Given that the list of conditioning variables x used in our framework is extensive and includes most of the important variables likely to affect bilateral trade flows, we feel that $\Gamma^* > 2$ is sufficiently large. In other words, we judge that any important omitted variables in ε are not likely to result in as much as an odds ratio of 2 in receiving treatment between the treated and untreated subject in any matched pair. Thus, we may accept these estimated treatment effects of 'Both in GATT/WTO' and 'GSP' with a reasonable degree of confidence.

When the matching criterion becomes more stringent such that further imaginable sources of selection bias are minimized, one may accept an even lower magnitude of Γ^*, as the remaining possibility of selection bias is lower. Both the 'matching within country-couple' and the 'matching within income-class combination' impose effective extra constraints relative to the 'unrestricted matching' benchmark, as discussed earlier and to some extent indicated by their fewer effective matched pairs. In the latter case of 'matching within income-class combination', the robustness of the treatment effect estimate is lower in general than the benchmark. The tolerance threshold (Γ^*) for positive selection bias now stands at 1.601, 1.807, and 1.391 (by the 80% caliper and the two-sided test) for the estimated treatment effect of 'Both in GATT/WTO', 'GSP', and 'One in GATT/WTO', respectively. It will take further investigations to make a fine judgement of whether the much lower thresholds of 1.601 and 1.391 are acceptable given the stricter criterion of matching within the same income-class combination. However, the treatment effect estimate of GSP may be regarded as reasonably robust with a tolerance level of $\Gamma^* = 1.807$ in the current stringent matching setup.

On the other hand, with the more stringent criterion of 'matching within country-couple', the robustness of the estimated treatment effects actually strengthens (as illustrated in Table 4). It now takes a strong presence of positive selection bias that leads to as much as an odds ratio of 2.543 (by the 80% caliper and the two-sided test), in receiving treatment between the treated and untreated subject in any matched pair, to nullify the original significance finding of 'Both in GATT/WTO' treatment effect. The corresponding figures are 2.494 and 1.772 for the 'GSP' and 'One in GATT/WTO' treatment, respectively. Thus, relative to the 'unrestricted matching' benchmark, it is even more comfortable for us to accept the results on the *Bothin* and the *GSP* treatment effects, as the degree of tolerance level for selection bias is higher despite that the possibility of remaining selection bias is lower with the extra matching criterion.

5.4 Overall Results

Overall, we conclude with two sets of relatively robust estimates for the trade effect of GATT/WTO membership, represented by the *Bothin* treatment effect obtained by 'unrestricted matching' and by 'matching within country-couple'. As discussed earlier, the results of 'unrestricted matching' are almost identical to those of 'matching within year' and 'matching within period' and capture most likely the cross-sectional treatment effect. Our intermediate estimate (by the 80% caliper) suggests that the GATT/WTO membership increases bilateral trade volume by 193% ($= e^{1.075} - 1$) for trading partners that are both GATT/WTO members, relative to trading partners that are both outside the GATT/WTO. On the other hand, the results of 'matching within country-couple' capture the time-series variation of treatment effect. In this case, our intermediate estimate (by the 80% caliper) suggests that the GATT/WTO membership increases bilateral trade volume by a smaller factor of 126% ($= e^{0.814} - 1$) when trading partners both join GATT/WTO, relative to when they both do not.

Regarding GSP, we are more comfortable with the results obtained based on 'matching within income-class combination', as the decision to extend GSP is very likely to be dependent on trading partners' relative income levels. The intermediate estimate (by the 80% caliper) suggests that the GSP scheme raises bilateral trade volume by a factor of 73% ($= e^{0.551} - 1$). Note again that the GATT/WTO effect has in the past applied to a much larger number of trading relationships (114,750) than the GSP effect (54,285). Thus, either on the average or in the aggregate, the realized trade promoting effect of GATT/WTO far exceeds that of GSP.

6. CONCLUSION

As the multilateral trade institution evolved from the simple trade treaty, the GATT, to the overarching organization, the WTO, it is timely to review whether the institution lives up to its objective of promoting international trade. In spite of the common affirmative impression, Rose (2004) argues that little evidence is found that the GATT/WTO has had an impact on trade. In this paper, we readdress this important policy question, arguing against the parametric gravity-equation approach used by Rose (2004), and propose nonparametric approaches to estimating the GATT/WTO trade effect. This involves using matching meth-

ods to estimate the 'treatment effect' of GATT/WTO membership and permutation-based inferential procedures to assess statistical significance of the estimated effects. The problem of potential selection bias is addressed by the Rosenbaum (2002) sensitivity analysis and by restricting matching within sub-samples.

Contrary to Rose (2004), we find consistent and large positive effects on trade when countries participate in the GATT/WTO system. The effect is larger when two countries in a bilateral trade relationship are both members than when only one is a member, but the effect is nonetheless significantly positive even when only one is a member, confirming an overall trade-creating effect of the multilateral institution. The positive effect of multilateral participation in the GATT/WTO system also exceeds that of the preferential GSP scheme by a great extent, either on the average or in the aggregate, contrasting Rose's (2004) result that GSP is more effective in promoting trade than the GATT/WTO. Overall, our finding confirms the common sense that preferential liberalizations (via GSP or unilateral participation in the GATT/WTO) may be good, but multilateral liberalizations (via reciprocal participation in the GATT/WTO) are most effective at promoting trade.

We conclude with some qualifications of our findings. First, the data set of Rose (2004) that we use includes only observations with positive bilateral trade flows. A recent study by Felbermayr and Kohler (2007) suggests that by taking into account trading relationships with zero trade leads to a bigger trade effect of GATT/WTO membership than indicated by Rose (2004). This finding does not undermine our foregoing conclusions, if we accept the notion proposed in Felbermayr and Kohler (2007) that non-existent trading relationships are also more likely those of non-participants in the GATT/WTO. It implies that if we already find significantly positive trade effect of GATT/WTO conditional on observations with positive trade flows, then the ultimate trade effect of GATT/WTO incorporating the extensive margin would only be larger. Second, using the data set of Rose (2004), which compiles the average of export and import flows, implies that we can not differentiate the direction of trade flows and thus control for exporter or importer specific effects. However, just as in Rose (2004), who controls for country-couple or dyad specific effect in some analysis with fixed-effect estimator, we accomplish the same by restricting matching to between observations of the same country-couple in one of our analysis. It may be helpful to further verify our findings by using directional trade data. We leave this for future research.

REFERENCES

Aakvik, A., 2001. Bounding a matching estimator: the case of a Norwegian training program. Oxford Bulletin of Economics and Statistics 63, 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, 151–184.

Anderson, J. E., 1979. A theoretical foundation for the gravity equation. American Economic Review 69 (1), 106–16.

Anderson, J. E., van Wincoop, E., 2003. Gravity with gravitas: A solution to the border puzzle. American Economic Review 93 (1), 170–92.

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, 281–325.

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.

Evenett, S. J., Braga, C. A. P., 2005. WTO accession: Lessons from experience. World Bank Trade Note 22.

Felbermayr, G., Kohler, W., 2007. Does WTO membership make a difference at the extensive margin of world trade? CESifo Working Paper No. 1898.

Good, P. I., 2004. Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3rd Edition. Springer Series in Statistics. Springer.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 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. The Review of Economic Studies 64 (4), 605–654.

Helpman, E., Krugman, P., 1985. Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy. The MIT Press, Cambridge, MA.

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

Hollander, M., Wolfe, D. A., 1999. Nonparametric statistical methods, 2nd Edition. Wiley.

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, 257–302.

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

Imbens, G. W., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. The 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 (Series A) 168, 109–126.

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. The Journal of Human Resources 35 (2), 347–375.

Lee, M.-J., 2005. Micro-econometrics for policy, program, and treatment effects. Oxford University Press.

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

Lehmann, E. L., Romano, J. P., 2005. Testing statistical hypotheses. 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.

Martiz, J., 1995. Distribution-free statistical methods, 2nd Edition. Chapman&Hall.

Oguledo, V. I., MacPhee, C. R., 1994. Gravity models: A reformulation and an application to discriminatory trade arrangements. Applied Economics 26 (2), 107–120.

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

Ramsey, J. B., 1969. Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society (Series B) 31 (2), 350–371.

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

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., 2005. Membership has its privileges: The impact of GATT on international trade. Stanford Center For International Development Working Paper No. 250.

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

WTO, 2006. International Trade Statistics. World Trade Organization, Geneva.

Table 1: misspecification test of the gravity equation

	(I) Rose		(II) RESET		(III) with interaction terms	
Both in GATT/WTO	-0.04	(0.05)	-0.04	(0.05)	-10.72	(1.10)
One in GATT/WTO	-0.06	(0.05)	-0.07	(0.05)	-7.61	(1.08)
GSP	0.86	(0.03)	0.99	(0.04)	0.56	(0.26)
Log distance	-1.12	(0.02)	-1.33	(0.05)	-1.10	(0.06)
Log product real GDP	0.92	(0.01)	1.08	(0.03)	0.86	(0.03)
Log product real GDP p/c	0.32	(0.01)	0.38	(0.02)	0.05	(0.04)
Regional FTA	1.20	(0.11)	1.55	(0.11)	0.58	(0.39)
Currency union	1.12	(0.12)	1.30	(0.13)	0.04	(0.32)
Common Language	0.31	(0.04)	0.35	(0.04)	0.09	(0.11)
Land border	0.53	(0.11)	0.67	(0.11)	0.56	(0.19)
Number landlocked	-0.27	(0.03)	-0.31	(0.03)	-0.17	(0.09)
Number islands	0.04	(0.04)	0.06	(0.04)	0.11	(0.12)
Log product land area	-0.10	(0.01)	-0.11	(0.01)	-0.17	(0.02)
Common colonizer	0.58	(0.07)	0.69	(0.07)	1.08	(0.16)
Currently colonized	1.08	(0.23)	1.28	(0.23)	4.81	(0.57)
Ever colony	1.16	(0.12)	1.45	(0.13)	-0.53	(0.21)
Common country	-0.02	(1.08)	-0.08	(1.07)	0.05	(1.03)
$(\widehat{\text{Log real trade}})^2$			-0.01	(0.002)		
Both in x GSP					0.11	(0.26)
Both in x Log distance					-0.03	(0.07)
Both in x Log product real GDP					0.07	(0.03)
Both in x Log product real GDP p/c					0.33	(0.05)
Both in x Regional FTA					0.27	(0.41)
Both in x Currency union					1.37	(0.35)
Both in x Common Language					0.30	(0.12)
Both in x Land border					0.02	(0.25)
Both in x Number landlocked					-0.13	(0.09)
Both in x Number islands					-0.10	(0.12)
Both in x Log product land area					0.10	(0.02)
Both in x Common colonizer					-0.61	(0.18)
Both in x Currently colonized					-3.91	(0.62)
Both in x Ever colony					1.69	(0.25)
Both in x Common country					(dropped)	
One in x GSP					0.48	(0.26)
One in x Log distance					-0.01	(0.06)
One in x Log product real GDP					0.05	(0.03)
One in x Log product real GDP p/c					0.25	(0.05)
One in x Regional FTA					1.17	(0.41)
One in x Currency union					0.67	(0.37)
One in x Common Language					0.27	(0.12)
One in x Land border					-0.08	(0.23)
One in x Number landlocked					-0.10	(0.09)
One in x Number islands					-0.10	(0.12)
One in x Log product land area					0.06	(0.02)
One in x Common colonizer					-0.58	(0.17)
One in x Currently colonized					(dropped)	
One in x Ever colony					1.71	(0.24)
One in x Common country					(dropped)	
Observations	234,597		234,597		234,597	
R^2	0.6480		0.6486		0.6525	
RMSE	1.9796		1.9777		1.9669	

Note:
Regressand: log real trade. OLS with year effects (intercepts not reported). Robust standard errors (clustering by country-couples) are in parentheses.

Table 2: trade effect based on gravity equation with interaction terms

	Both in Effect	One in Effect
Percentiles		
Smallest	-3.2530	-2.5517
25%	-0.2579	-0.2096
50%	0.2552	0.2093
75%	0.7786	0.7118
Largest	4.8556	4.8237
Mean	0.2724	0.2723
Standard Error of Mean	0.0017	0.0015
Observations	234,597	234,597
Standard Deviation	0.8428	0.7442
Variance	0.7103	0.5538
Skewness	0.3219	0.6496
Kurtosis	3.9576	4.4601

Note:
Calculation based on regression (III) of Table 1. For trading partners i and j in year t,
(i) $\{\text{Both in effect}\}_{ijt} = \beta_{\{\text{Both in}\}} + \beta_{\{\text{Both in x GSP}\}} \{\text{GSP}\}_{ijt} + \ldots + \beta_{\{\text{Both in x Ever colony}\}} \{\text{Ever colony}\}_{ijt}$;
(ii) $\{\text{One in effect}\}_{ijt} = \beta_{\{\text{One in}\}} + \beta_{\{\text{One in x GSP}\}} \{\text{GSP}\}_{ijt} + \ldots + \beta_{\{\text{One in x Common colonizer}\}} \{\text{Common colonizer}\}_{ijt} + \beta_{\{\text{One in x Ever colony}\}} \{\text{Ever colony}\}_{ijt}$.

Table 3: treatment effect – unrestricted matching

	permutation test				signed-rank test				sensitivity analysis			
									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^+
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^+
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^+
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^+
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^+
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^+
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^+

Note:
1. The pool of potential matches for an observation are observations with the opposite treatment; no further restriction is imposed. The numbers of matched pairs obtained (M_1 for the case of the effect on the treated, and M_0 for the untreated) are specified in the parentheses.
2. The caliper choice is set such that only 100%, 80%, or 60% of matched pairs are qualified for the estimation of the treatment effect. For example, for the case of 60%, matched pairs with a distance in terms of x exceeding the upper 60 percentile of all matched pairs are discarded.
3. The regressors in the benchmark gravity equation of Rose (2004) are used as the conditioning variables x for matching, with the treatment dummy variable being investigated (*Bothin*, *Onein*, or *GSP*) suitably excluded from the list.
4. In the column 'permutation test', the 'effect' sub-column presents the treatment effect estimate based on the D statistic; the p-value is obtained for the observed D statistic using the permutation test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.
5. In the column 'signed-rank test', the 'effect' sub-column presents the treatment effect estimate based on the Hodges and Lehmann (1963) estimator; the p-value is obtained for the observed R statistic using the signed-rank test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.
6. In the column 'sensitivity analysis', the sensitivity analysis is conducted for the above signed-rank test based on a significance level of $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of Γ) indicates the relevant distribution on which the bound for the p-value of the signed-rank test is based. Γ^* indicates the critical value of Γ at which the conclusion of the signed-rank test reverses.

Table 4: treatment effect – matching within country-couple

	permutation test			signed-rank test			sensitivity analysis			
							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 ($M_1 = 19,760$):										
100%	0.941	0.000	[0.912, 0.970]	0.996	0.000	[0.970, 1.023]	3.189	R^+	3.170	R^+
80%	0.760	0.000	[0.727, 0.792]	0.814	0.000	[0.785, 0.844]	2.559	R^+	2.543	R^+
60%	0.833	0.000	[0.797, 0.870]	0.876	0.000	[0.843, 0.910]	2.792	R^+	2.771	R^+
on the untreated ($M_0 = 9,510$):										
100%	1.300	0.000	[1.255, 1.345]	1.359	0.000	[1.317, 1.401]	4.168	R^+	4.129	R^+
80%	1.117	0.000	[1.067, 1.167]	1.161	0.000	[1.115, 1.207]	3.475	R^+	3.440	R^+
60%	0.989	0.000	[0.931, 1.048]	1.019	0.000	[0.967, 1.071]	3.017	R^+	2.983	R^+
on all ($M_1 + M_0 = 29,270$):										
100%	1.058	0.000	[1.033, 1.082]	1.110	0.000	[1.087, 1.132]	3.515	R^+	3.496	R^+
80%	0.895	0.000	[0.868, 0.923]	0.943	0.000	[0.918, 0.967]	2.927	R^+	2.911	R^+
60%	0.935	0.000	[0.903, 0.967]	0.969	0.000	[0.940, 0.998]	3.021	R^+	3.002	R^+
One in GATT/WTO treatment effect										
on the treated ($M_1 = 23,463$):										
100%	0.464	0.000	[0.438, 0.489]	0.532	0.000	[0.510, 0.555]	1.940	R^+	1.931	R^+
80%	0.403	0.000	[0.374, 0.432]	0.470	0.000	[0.444, 0.495]	1.782	R^+	1.772	R^+
60%	0.371	0.000	[0.335, 0.407]	0.460	0.000	[0.428, 0.491]	1.666	R^+	1.656	R^+
on the untreated ($M_0 = 15,182$):										
100%	0.579	0.000	[0.544, 0.613]	0.641	0.000	[0.611, 0.671]	2.110	R^+	2.097	R^+
80%	0.463	0.000	[0.423, 0.503]	0.525	0.000	[0.492, 0.559]	1.818	R^+	1.805	R^+
60%	0.386	0.000	[0.339, 0.432]	0.430	0.000	[0.390, 0.469]	1.610	R^+	1.597	R^+
on all ($M_1 + M_0 = 38,645$):										
100%	0.509	0.000	[0.488, 0.529]	0.574	0.000	[0.556, 0.592]	2.023	R^+	2.016	R^+
80%	0.428	0.000	[0.404, 0.452]	0.492	0.000	[0.472, 0.513]	1.816	R^+	1.808	R^+
60%	0.403	0.000	[0.374, 0.432]	0.478	0.000	[0.453, 0.503]	1.706	R^+	1.698	R^+
GSP treatment effect										
on the treated ($M_1 = 52,025$):										
100%	0.487	0.000	[0.476, 0.499]	0.478	0.000	[0.468, 0.488]	2.579	R^+	2.570	R^+
80%	0.492	0.000	[0.479, 0.506]	0.489	0.000	[0.478, 0.501]	2.504	R^+	2.494	R^+
60%	0.379	0.000	[0.363, 0.395]	0.371	0.000	[0.357, 0.385]	1.945	R^+	1.937	R^+

Note:
1. The pool of potential matches for an observation are restricted to observations with the opposite treatment and of the same country-couple; observations without a match are discarded. The numbers of matched pairs obtained (M_1 for the case of the effect on the treated, and M_0 for the untreated) are specified in the parentheses.
2. The caliper choice is set such that only 100%, 80%, or 60% of matched pairs are qualified for the estimation of the treatment effect. For example, for the case of 60%, matched pairs with a distance in terms of x exceeding the upper 60 percentile of all matched pairs are discarded.
3. The regressors in the benchmark gravity equation of Rose (2004) are used as the conditioning variables x for matching, with the treatment dummy variable being investigated (*Bothin*, *Onein*, or *GSP*) suitably excluded from the list.
4. In the column 'permutation test', the 'effect' sub-column presents the treatment effect estimate based on the D statistic; the p-value is obtained for the observed D statistic using the permutation test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.
5. In the column 'signed-rank test', the 'effect' sub-column presents the treatment effect estimate based on the Hodges and Lehmann (1963) estimator; the p-value is obtained for the observed R statistic using the signed-rank test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.
6. In the column 'sensitivity analysis', the sensitivity analysis is conducted for the above signed-rank test based on a significance level of $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of Γ) indicates the relevant distribution on which the bound for the p-value of the signed-rank test is based. Γ^* indicates the critical value of Γ at which the conclusion of the signed-rank test reverses.

Table 5: treatment effect – matching within year

	permutation test				signed-rank test				sensitivity analysis			
									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%	1.329	0.000	[1.308,	1.350]	1.334	0.000	[1.314,	1.354]	2.433	R^+	2.427	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^+
on the untreated ($M_0 = 21,037$):												
100%	0.340	0.000	[0.298,	0.381]	0.305	0.000	[0.267,	0.344]	1.251	R^+	1.245	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^+
on all ($M_1 + M_0 = 135,787$):												
100%	1.176	0.000	[1.157,	1.194]	1.164	0.000	[1.146,	1.181]	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^+
One in GATT/WTO treatment effect												
on the treated ($M_1 = 98,810$):												
100%	0.761	0.000	[0.739,	0.782]	0.767	0.000	[0.748,	0.786]	1.751	R^+	1.747	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^+
on the untreated ($M_0 = 21,037$):												
100%	0.032	0.054	[-0.007,	0.072]	0.038	0.014	[0.003,	0.071]	1.009	R^+	1.004	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^+
on all ($M_1 + M_0 = 119,847$):												
100%	0.633	0.000	[0.614,	0.652]	0.628	0.000	[0.611,	0.645]	1.605	R^+	1.601	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^+
GSP treatment effect												
on the treated ($M_1 = 54,285$):												
100%	0.850	0.000	[0.830,	0.870]	0.791	0.000	[0.773,	0.810]	2.275	R^+	2.267	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^+

Note:

1. The pool of potential matches for an observation are restricted to observations with the opposite treatment and of the same year; observations without a match are discarded. The numbers of matched pairs obtained (M_1 for the case of the effect on the treated, and M_0 for the untreated) are specified in the parentheses.

2. The caliper choice is set such that only 100%, 80%, or 60% of matched pairs are qualified for the estimation of the treatment effect. For example, for the case of 60%, matched pairs with a distance in terms of x exceeding the upper 60 percentile of all matched pairs are discarded.

3. The regressors in the benchmark gravity equation of Rose (2004) are used as the conditioning variables x for matching, with the year dummies and the treatment dummy variable being investigated (*Bothin*, *Onein*, or *GSP*) suitably excluded from the list.

4. In the column 'permutation test', the 'effect' sub-column presents the treatment effect estimate based on the D statistic; the p-value is obtained for the observed D statistic using the permutation test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.

5. In the column 'signed-rank test', the 'effect' sub-column presents the treatment effect estimate based on the Hodges and Lehmann (1963) estimator; the p-value is obtained for the observed R statistic using the signed-rank test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.

6. In the column 'sensitivity analysis', the sensitivity analysis is conducted for the above signed-rank test based on a significance level of $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of Γ) indicates the relevant distribution on which the bound for the p-value of the signed-rank test is based. Γ^* indicates the critical value of Γ at which the conclusion of the signed-rank test reverses.

Table 6: treatment effect – matching within period

	permutation test				signed-rank test				sensitivity analysis			
									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%	1.331	0.000	[1.310,	1.352]	1.336	0.000	[1.316,	1.356]	2.438	R^+	2.432	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^+
on the untreated ($M_0 = 21,037$):												
100%	0.340	0.000	[0.298,	0.381]	0.305	0.000	[0.267,	0.344]	1.251	R^+	1.245	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^+
on all ($M_1 + M_0 = 135,787$):												
100%	1.177	0.000	[1.158,	1.196]	1.165	0.000	[1.147,	1.183]	2.213	R^+	2.208	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^+
One in GATT/WTO treatment effect												
on the treated ($M_1 = 98,810$):												
100%	0.762	0.000	[0.741,	0.784]	0.768	0.000	[0.749,	0.787]	1.753	R^+	1.749	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^+
on the untreated ($M_0 = 21,037$):												
100%	0.032	0.054	[-0.007,	0.072]	0.038	0.015	[0.003,	0.071]	1.009	R^+	1.004	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^+
on all ($M_1 + M_0 = 119,847$):												
100%	0.634	0.000	[0.615,	0.653]	0.629	0.000	[0.612,	0.646]	1.606	R^+	1.603	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^+
GSP treatment effect												
on the treated ($M_1 = 54,285$):												
100%	0.851	0.000	[0.831,	0.871]	0.792	0.000	[0.773,	0.811]	2.276	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^+

Note:

1. The pool of potential matches for an observation are restricted to observations with the opposite treatment and of the same period, where the periods are: 1948 (Before Annecy round), 1949-1951 (Annecy to Torquay round), 1952-1956 (Torquay to Geneva round), 1957-1961 (Geneva to Dillon round), 1962-1967 (Dillon to Kennedy round), 1968-1979 (Kennedy to Tokyo round), 1980-1994 (Tokyo to Uruguay round), 1995-(After Uruguay round). Observations without a match are discarded. The numbers of matched pairs obtained (M_1 for the case of the effect on the treated, and M_0 for the untreated) are specified in the parentheses.

2. The caliper choice is set such that only 100%, 80%, or 60% of matched pairs are qualified for the estimation of the treatment effect. For example, for the case of 60%, matched pairs with a distance in terms of x exceeding the upper 60 percentile of all matched pairs are discarded.

3. The regressors in the benchmark gravity equation of Rose (2004) are used as the conditioning variables x for matching, with the treatment dummy variable being investigated (*Bothin*, *Onein*, or *GSP*) suitably excluded from the list.

4. In the column 'permutation test', the 'effect' sub-column presents the treatment effect estimate based on the D statistic; the p-value is obtained for the observed D statistic using the permutation test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.

5. In the column 'signed-rank test', the 'effect' sub-column presents the treatment effect estimate based on the Hodges and Lehmann (1963) estimator; the p-value is obtained for the observed R statistic using the signed-rank test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.

6. In the column 'sensitivity analysis', the sensitivity analysis is conducted for the above signed-rank test based on a significance level of $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of Γ) indicates the relevant distribution on which the bound for the p-value of the signed-rank test is based. Γ^* indicates the critical value of Γ at which the conclusion of the signed-rank test reverses.

Table 7: treatment effect – matching within income-class combination

	permutation test				signed-rank test				sensitivity analysis			
									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 ($M_1 = 112,959$):												
100%	1.124	0.000	[1.103,	1.146]	1.097	0.000	[1.076,	1.118]	2.024	R^+	2.019	R^+
80%	0.778	0.000	[0.753,	0.802]	0.734	0.000	[0.711,	0.757]	1.605	R^+	1.601	R^+
60%	0.541	0.000	[0.514,	0.569]	0.504	0.000	[0.478,	0.530]	1.389	R^+	1.385	R^+
on the untreated ($M_0 = 21,013$):												
100%	0.309	0.000	[0.268,	0.351]	0.274	0.000	[0.236,	0.312]	1.222	R^+	1.216	R^+
80%	0.175	0.000	[0.128,	0.222]	0.131	0.000	[0.089,	0.173]	1.083	R^+	1.077	R^+
60%	0.101	0.000	[0.047,	0.155]	0.059	0.008	[0.010,	0.107]	1.015	R^+	1.009	R^+
on all ($M_1 + M_0 = 133,972$):												
100%	0.997	0.000	[0.977,	1.016]	0.955	0.000	[0.936,	0.973]	1.883	R^+	1.880	R^+
80%	0.662	0.000	[0.640,	0.684]	0.612	0.000	[0.592,	0.633]	1.507	R^+	1.504	R^+
60%	0.442	0.000	[0.418,	0.467]	0.402	0.000	[0.379,	0.424]	1.313	R^+	1.309	R^+
One in GATT/WTO treatment effect												
on the treated ($M_1 = 98,363$):												
100%	0.650	0.000	[0.627,	0.672]	0.628	0.000	[0.608,	0.648]	1.556	R^+	1.552	R^+
80%	0.476	0.000	[0.452,	0.500]	0.460	0.000	[0.438,	0.481]	1.394	R^+	1.391	R^+
60%	0.342	0.000	[0.315,	0.370]	0.324	0.000	[0.300,	0.347]	1.267	R^+	1.263	R^+
on the untreated ($M_0 = 21,013$):												
100%	0.049	0.007	[0.010,	0.087]	0.034	0.023	[0.000,	0.067]	1.006	R^+	1.001	R^+
80%	0.063	0.002	[0.020,	0.105]	0.051	0.003	[0.014,	0.087]	1.020	R^+	1.014	R^+
60%	0.034	0.084	[-0.014,	0.083]	0.028	0.092	[-0.012,	0.070]	1.007	R^-	1.013	R^-
on all ($M_1 + M_0 = 119,376$):												
100%	0.544	0.000	[0.524,	0.563]	0.512	0.000	[0.495,	0.530]	1.455	R^+	1.452	R^+
80%	0.391	0.000	[0.369,	0.412]	0.369	0.000	[0.350,	0.388]	1.320	R^+	1.316	R^+
60%	0.215	0.000	[0.192,	0.239]	0.200	0.000	[0.179,	0.219]	1.164	R^+	1.161	R^+
GSP treatment effect												
on the treated ($M_1 = 53,811$):												
100%	0.732	0.000	[0.712,	0.752]	0.693	0.000	[0.674,	0.712]	2.018	R^+	2.011	R^+
80%	0.588	0.000	[0.567,	0.608]	0.551	0.000	[0.531,	0.570]	1.813	R^+	1.807	R^+
60%	0.507	0.000	[0.485,	0.530]	0.474	0.000	[0.452,	0.496]	1.706	R^+	1.699	R^+

Note:

1. The pool of potential matches for an observation are restricted to observations with the opposite treatment and of the same income-class combination, where the income-class combinations are: 'low income-low income' country-couples, 'low income-middle income' country couples, 'low income-high income' country couples, 'middle income-middle income' country couples, 'middle income-high income' country couples, and 'high income-high income' country couples. Observations without a match are discarded. The numbers of matched pairs obtained (M_1 for the case of the effect on the treated, and M_0 for the untreated) are specified in the parentheses.
2. The caliper choice is set such that only 100%, 80%, or 60% of matched pairs are qualified for the estimation of the treatment effect. For example, for the case of 60%, matched pairs with a distance in terms of x exceeding the upper 60 percentile of all matched pairs are discarded.
3. The regressors in the benchmark gravity equation of Rose (2004) are used as the conditioning variables x for matching, with the treatment dummy variable being investigated (*Bothin*, *Onein*, or *GSP*) suitably excluded from the list.
4. In the column 'permutation test', the 'effect' sub-column presents the treatment effect estimate based on the D statistic; the p-value is obtained for the observed D statistic using the permutation test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.
5. In the column 'signed-rank test', the 'effect' sub-column presents the treatment effect estimate based on the Hodges and Lehmann (1963) estimator; the p-value is obtained for the observed R statistic using the signed-rank test based on the normal approximation approach; the CI is obtained by inverting the test procedure as discussed in the main text.
6. In the column 'sensitivity analysis', the sensitivity analysis is conducted for the above signed-rank test based on a significance level of $\alpha = 0.05$ in a one-sided or two-sided test. R^+ or R^- (as a function of Γ) indicates the relevant distribution on which the bound for the p-value of the signed-rank test is based. Γ^* indicates the critical value of Γ at which the conclusion of the signed-rank test reverses.