Review of International Economics, 18(1), 63–80, 2010
DOI:10.1111/j.1467-9396.2009.00877.x

Reciprocal Trade Agreements in Gravity Models: A Meta-Analysis

*Maria Cipollina and Luca Salvatici**

Abstract

The gravity model is a workhorse tool applicable in a wide range of empirical fields. It is regularly used to estimate the impact of reciprocal trade agreements (RTAs) on trade flows between partners. The studies report very different estimates since there is a significant difference in datasets, sample sizes, and independent variables. This paper combines, explains, and summarizes a large number of results using a meta-analysis approach. We provide pooled estimates, obtained from fixed and random effects models of the RTAs' effect size on bilateral trade: the hypothesis that there is no effect of RTAs on trade is robustly rejected at standard significance levels. The information collected on each estimate allows us to test the sensitivity of the results to alternative specifications and differences in the control variables considered, as well as the impact of the publication selection process.

1. Introduction

Preferential agreements are discriminatory policies related to trade liberalization with respect to a subset of trading partners. The world's trading system is characterized by a wide variety of preferential agreements which can be categorized into two broad types: reciprocal (bilateral) involving symmetric trade liberalization, and nonreciprocal (unilateral) involving asymmetric trade liberalization which provides countries with improved market access without opening up their domestic markets. The configuration of these agreements is diverse and is increasing in complexity, with overlapping agreements within and across continents, resulting in what Bhagwati et al. (1999) call a "spaghetti bowl" of trade relationships. The world has witnessed a veritable explosion of reciprocal preferential trade agreements since the mid-1990s: of the 250 agreements notified to the General Agreement on Tariffs and Trade and the World Trade Organization between 1947 and 2002, about half were notified after 1995. More than half of world trade now occurs within actual or prospective trading blocs, and nearly every country in the world is a signatory of one or more agreements (Lawrence, 1996; Clarete et al., 2003).

Given the multitude of reciprocal trade agreements (RTAs) in existence, and the political and economic resources devoted by world governments to forging these agreements, it is not surprising that there is a vast literature on the economic impacts of

* Cipollina: University of Molise, via De Sanctis, 86100 Campobasso, Italy. Tel: +39-0874-4041; Fax: +39-0874-311124; E-mail: cipollina@unimol.it. Salvatici: University of Molise, via De Sanctis, 86100 Campobasso, Italy. Tel: +39-0874-4041; Fax: +39-0874-311124; E-mail: luca.salvatici@unimol.it. We acknowledge financial support from the "Agricultural Trade Agreements (TRADEAG)" (Specific Targeted Research Project, contract 513666) and the "New Issues in Agricultural, Food and Bio-Energy Trade (AGFOODTRADE)" (Small and Medium-Scale Focused Research Project, grant agreement 212036) research projects, funded by the European Commission. We thank an anonymous referee for helpful comments.

© 2010 Blackwell Publishing Ltd

RTAs. This paper focuses on estimating their effects on trade: it is expected that RTAs increase trade between partners, since cheaper imports facilitated by the agreement replaces domestic production (trade creation) or crowds out imports from the rest of the world (trade diversion) (Viner, 1950; Meade, 1955). The influence of regional agreements on trade patterns has been estimated primarily through gravity equations assessing the differences between potential and actual trade flows (Baldwin, 1994; Eichengreen and Irwin, 1996; Feenstra, 1998; Anderson and van Wincoop, 2003). There have been numerous studies since the early 1970s, which was the period when the gravity equation was applied for the first time to international trade. The standard formulation of the gravity equation expresses bilateral trade between country *i* and country *j* as:

$$\ln T_{ij} = \beta_0 + \beta_1 \ln(Y_i) + \beta_2 \ln(Y_j) + \beta_3 \ln(Dist_{ij}) + \beta_4 Adj_{ij} + \beta_5 Lang_{ij} + \gamma RTA_{ij} + \varepsilon_{ij}, \quad (1)$$

where T_{ij} is the country pair's trade flow; $Y_{i(j)}$ indicates GDP or GNP of *i* and *j*; $Dist_{ij}$ is the distance between *i* and *j*; Adj_{ij}, $Lang_{ij}$, and RTA_{ij} are binary variables for common land border, language and reciprocal trade agreements, respectively; and ε_{ij} is the error term.

The results of these studies show disconcerting variance: the coefficients of *RTA* are not stable, with widely varying estimates across studies and some worrying rankings of trade-creating agreements. The World Bank (2005) provides a meta-analysis of the literature on the impact of RTAs on intra- and extra-regional trade. The analysis considers 17 research studies providing 362 estimates of the impact on the level of trade between partners. The mean value of these estimates is positive, but there is a high degree of variance around the mean.

We examine why *ex post* measurements of the (average) trade impact of RTAs appears so volatile. The question addressed in this paper is how much are these results influenced by flaws in the standard gravity equation estimation in terms of econometric specification and proper identification of the impact of the RTAs. To answer this we use meta-analysis (MA) to summarize and analyze the trade effects highlighted in the literature.

MA is a methodology for reviewing the literature, not an alternative approach to studying the trade effects of RTAs. The goal is not to discover the "true" value of the parameter under investigation, but rather to explain why there is so much variation in the empirical results reported in the economic studies that supposedly investigate the same phenomenon. Regression analysis of the existing regression analyses represents a methodology for quantitatively combining all these estimates (commonly referred to as the "effect size"), investigating their sensitivity to variations in the underlying assumptions, identifying and filtering out possible biases, and explaining the diversity in the results of these studies in terms of heterogeneity of study features (Rose and Stanley, 2005).

In this paper, multiple estimates from a single study are considered, to test for correlation within and between studies. We use point estimates of the relevant parameters from various studies as the individual observations for the meta-regression analysis (MRA) models, adopting weighted least squares (WLS) and checking the robustness and sensitivity of our results. We then focus on the effect of specific trade agreements on bilateral trade. Finally, we run a probit regression in order to identify which factors account for the positive and significant impact of RTAs on bilateral trade flows.

© 2010 Blackwell Publishing Ltd

2. Methodological Issues

MA is a set of quantitative techniques for evaluating and combining the empirical results from a group of studies. The main focus of MA is to test the null hypothesis that different point estimates, when treated as individual observations (γ), are equal to zero when the findings from this entire area of research are combined. MA has become increasingly popular in economics: in 2005, the *Journal of Economic Surveys* devoted a special issue (Vol. 19, No. 3) to this approach.

The first step in an MA, namely constructing a database of estimates, is crucial. Here, we drew on English-language papers, selected via extensive Google searches and searches of databases such as EconLit and the Web of Science. We also identified some papers cross-referenced in other works.

We searched on keywords in title, abstract, or subject: "trade agreements," "gravity equation," and "gravity model." The first identified papers dealing with trade agreements; the other two identified papers using gravity as their approach. From the first group of papers, we selected papers analyzing trade agreements that focused on bilateral trade flows; from the second group we selected studies that included trade agreements as control variables in the gravity equations. Our sample includes 85 papers (38 published academic journal articles, 47 working papers or unpublished studies) providing 1827 point estimates of the impact of RTAs on bilateral trade, i.e. the coefficient γ in equation (1).[2] Since some agreements change from "unilateral" to "reciprocal" over time, we do not consider estimates referring to periods when there were only preferential tariff reductions.

The choice to limit our review of the impact of RTAs to gravity models could be challenged on the grounds that some of the literature (e.g. Ghosh and Yamarik, 2004) argues that the pervasive trade creation effect of most RTAs is picking up a fragile relationship generated by an individual researcher's specification of the gravity model equation.[1] Although MA is an attractive technique for evaluating and combining empirical results, analyzing completely different methodologies can be risky (the apples and oranges problem referred to by Glass et al., 1981). We do not claim to establish professional consensus or identify a clear and uncontroversial approach to the evaluation of the impact of RTAs using MA. Our more modest goal is to provide an assessment of the methodological choices and possible (relative) biases induced by different model specifications within this large and growing literature using the gravity approach.

One of the main criticisms of MA is that because the quality of studies included in the dataset can vary considerably, strong methodological or empirical analyses are lumped together with studies that may have serious methodological or empirical limitations (the "garbage in, garbage out" criticism, as our anonymous referee phrased it). It is argued that alternative selection schemes might be considered arbitrary and subjective. However, the more substantial reason proposed for the inclusion in meta-databases of both published and unpublished studies, is to reduce the so-called "publication bias" (Ashenfelter et al., 1999). In what follows, we show that some interesting results can be obtained from using a broader database, despite the conventional wisdom on how this "bias" should be interpreted, with which we do not completely agree.

A single study quite often can enable multiple estimates of the effect being considered. The presence of more than one estimate per study is problematic because the assumption that multiple observations from the same study are independent is too strong. On the other hand, counting all estimates equally would tend to overburden

© 2010 Blackwell Publishing Ltd

studies with many estimates (Stanley, 2001). Various solutions have been suggested in the literature. Some authors include a dummy variable (fixed effect) for each study providing more than one observation (Jarrell and Stanley, 1990), others use a panel specification (Jeppesen et al., 2002; Disdier and Head, 2004), and a few scholars have attempted to incorporate the hierarchical structure of the data in multilevel linear models (e.g. de Dominicis et al., 2006). Alternatively, each study can be represented with a single observation, identifying a "preferred" estimate, using averages or medians of the estimates from each paper or randomly selecting one estimate (Card and Krueger, 1995; Stanley, 2001; Rose and Stanley, 2005): in this case, important information is lost in the grouping process and it is not clear which estimate would be the most useful (Jeppesen et al., 2002).

Pooling different estimates into a large sample for MA raises the question of within-study versus between-study heterogeneity. In order to account for this, we can distinguish between a fixed effects model (FEM) and a random effects model (REM). The FEM assumes that differences across studies are due only to within-variation. Following Higgins and Thompson (2002), the single "true" effect ($\hat{\theta}_F$), underlying every study, is calculated as the weighted average of the study estimates, using the precisions as weights:

$$\hat{\theta}_F = \frac{\sum_{i=1}^{n} \hat{\theta}_i w_i}{\sum_{i=1}^{n} w_i}, \tag{2}$$

where $\hat{\theta}_i$ is the individual estimate of the effect of RTAs (our γ_i) and the weight w_i is inversely proportional to the square of its standard error ($Se(\hat{\theta}_i)$), so that studies with smaller standard errors have greater weight than studies with larger standard errors.

A field of the literature showing high heterogeneity cannot be summarized by the fixed effects estimate under the assumption that a single "true" effect underlies every study. Consequently, the fixed effects estimator is inconsistent and the REM is more appropriate. The REM considers both between-study and within-study variability, and assumes that the studies are a random sample from the universe of all possible studies (Sutton et al., 2000). Unlike in the FEM, the individual studies are not assumed to be estimating a true single effect size, but the true effects in each study are assumed to have been sampled from a distribution of effects under a normal distribution with a mean of 0 and variance of τ^2. The weights, w_i, incorporate an estimate of between-study heterogeneity, $\hat{\tau}^2$, and are equal to $(w_i^{-1} + \tau^2)^{-1}$.

Allowing for between-study variation has the effect of reducing the relative weighting given to the more precise studies. Hence, the REM produces a more conservative confidence interval for the pooled effects estimate. A test of homogeneity of the θ_i is provided by referring the statistic $Q = \sum_{i=1}^{n} w\left(\hat{\theta}_i - \hat{\theta}_F\right)^2$ to a χ^2-distribution with $n - 1$ degrees of freedom. If Q exceeds the upper-tail critical value, the observed variance in estimated effect sizes is greater than what we would expect to occur by chance if all studies shared the same "true" parameter (Higgins and Thompson, 2002).[3] The Q-test should be used with caution because its power is low (Florax, 2002): when the sample of observations is large, for example, homogeneity is likely to be rejected even when the individual effect sizes differ only slightly. In any case, its computation is an intermediate step in computing the preferred tests (H^2 and I^2) that we use in our analysis. The statistic H^2 provides a possible measure of the degree of heterogeneity, through the ratio of Q over its degrees of freedom. Homogeneity in effect sizes is indicated by the statistic H^2

© 2010 Blackwell Publishing Ltd

being close to 1. The I^2 statistic, on the other hand, measures the percentage of variability in point estimates that is due to heterogeneity rather than sampling error:

$$I^2 = \frac{H^2 - 1}{H^2} = \frac{Q - n + 1}{Q}. \tag{3}$$

In this case, then, a value close to 100 would support the assumption of heterogeneity.

The simple mean of estimates could be misleading in the presence of more than one mode or outlier in the sample of estimates because a large part of the estimates may lie to one side of the mean value. If the distribution is multimodal or if there are outliers (extreme data points), the mean could be biased. Skewness is usually tested by comparing the mode, median, and mean of the distribution. However, this would not be accurate for symmetrically distributed outliers, since they tend to cancel each other out, or when outliers have smaller statistical weights than other data points so that they contribute less to the mean. Some authors prefer to remove outliers since they compress the variation in the rest of the sample and are likely to lead to "fragile" findings (Disdier and Head, 2004), while others claim that removing outliers and extreme results at an early stage in the MA could introduce substantial bias in the meta-results, and the influence of removing outliers should be explored in a sensitivity analysis (Stanley, 2001; Florax, 2002).

Finally, we have referred to the general belief that publication bias occurs when researchers, referees, or editors have a preference for statistically significant results.[4] Several meta-regression and graphical methods have been proposed to differentiate genuine empirical effects from publication bias.

The simplest and most conventional method of detecting publication impact is inspection of a funnel graph diagram. This is a scatter diagram presenting a measure of sample size or precision of the estimate, as the inverse of the standard error ($1/Se$), on the vertical axis, and the measured effect size on the horizontal axis. Asymmetry is the mark of publication impact: in the absence of publication impact, the estimates will vary randomly and symmetrically around the true effect (Stanley, 2005). The diagram should therefore resemble an inverted funnel, wide at the bottom for small-sample studies and narrowing as it rises.

A meta-regression analysis (MRA) model can also be used to investigate and account for publication impact. The model regresses estimated coefficients (γ_i) on their standard errors (Card and Krueger, 1995; Ashenfelter et al., 1999):

$$\gamma_i = \beta_1 + \beta_0 Se_i + \varepsilon_i. \tag{4}$$

In the absence of publication selection, the magnitude of the reported effect will vary randomly around the value β_1, independent of its standard error.

Although it is true that the peer-review process can greatly affect the magnitude of the estimated effect, whether or not this impact should be considered a bias is a moot point. Since in MA, notwithstanding the wide variation in the quality of the point estimates included in the study, each estimate in the sample is weighted equally; it could be argued that there is a *nonpublication bias* due to the lower quality of unpublished research. In the following we assess the consequences of the peer-review process, but refer to a general "publication impact" (rather than a "bias") for the above reasons.

© 2010 Blackwell Publishing Ltd

3. MA Regression

The standard meta regression model includes a set of explanatory variables (X) to integrate and explain the diverse findings in the literature. Since the studies in the literature can differ greatly in terms of datasets, sample sizes, and independent variables, the variance of the estimated coefficients may not be equal. As a result, meta-regression errors are likely to be heteroskedastic, although the ordinary least squares (OLS) estimates of the MRA coefficients remain unbiased and consistent. A WLS corrects the MRA for heteroskedasticity and allows efficient estimates with correct standard errors. The WLS version is obtained by dividing regression equation (4) by the individual estimated standard errors (Stanley and Jarrell, 2005). The potential for heteroskedasticity, therefore, directs the meta-analyst's attention to the reported t-statistics ($t_i = \gamma_i/Se_i$):

$$\frac{\gamma_{ji}}{Se_{ji}} = t_{ji} = \beta_0 + \beta_1\left(\frac{1}{Se_{ji}}\right) + \sum_{k=1}^{K}\left(\frac{\alpha_k X_{jik}}{Se_{ji}}\right) + e_i, \tag{5}$$

where γ_{ji} is the reported estimate i of the jth study in the literature, β expresses the true value of the parameter of interest, X_{jik} is the independent variable which measures the relevant characteristics in an empirical study and explains its systematic variation from other results in the literature, α_k is the regression coefficient which reflects the biasing effect of particular study characteristics, and e_i is the disturbance term.

This regression may still lead to inefficient, though consistent, estimators since it does not consider the dependence of estimates obtained in the same study. In order to obtain correct standard errors, we adopt a "robust with cluster" procedure, adjusting standard errors for intra-study correlation. Each cluster identifies the study to which the estimate belongs: this changes the variance–covariance matrix and the standard errors of the estimators, but not the estimated coefficients themselves.

Finally, we adopt a specification that investigates the factors influencing whether the estimated effects are positive and significantly different from zero. The estimated model is given by:

$$s_{ji} = a + \sum_{k=1}^{K} b_k X_{jik} + e_{ji}, \tag{6}$$

where the dependent variable is a dummy that takes the value 1 if the estimated effect size is positive and statistically significant. The probability that an estimated effect size is positive and significant is explained by a set of explanatory variables (X) and is estimated by running a probit regression.

Explanatory Variables

Our set of variables X in equation (5) can be divided into two groups: the first includes dummies explaining the diversity in the results from a methodological point of view, and the second includes dummies regarding the features of the studies considered. The methodological dummies included in the MRA are based on a recent survey of the errors in the empirical literature applying gravity equations, carried out by Baldwin and Taglioni (2006). They rank the major errors by assigning different "medals" according to the seriousness of the implied consequences.

© 2010 Blackwell Publishing Ltd

The *gold medal* classic gravity model mistakes arise from the correlation between the omitted variables and the trade-cost terms, which leads to biased estimates. In particular, the estimated trade impact will be upward-biased if the omitted variables and the variable of interest (RTAs, in our case) are positively correlated. "The point is that the formation of currency unions is not random but rather driven by many factors, including many of the factors omitted from the gravity regression" (Baldwin and Taglioni, 2006, p. 9): apparently, the same point can be made for RTAs.

Possible solutions to the gold medal problem include country effects (one dummy for all trade flows that involve a particular country) and pair effects (one dummy for all observations of trade between a given pair of countries). Country dummies remove the cross-section but not the time-series bias, and this is a serious shortcoming since omitted factors affecting bilateral trade costs often vary over time. Pair dummies can only be used with panel data (in a cross-section analysis the number of dummies would be equal to the number of observations) and, in any case, they provide a partial answer to the gold medal bias (Baldwin and Taglioni, 2006).

Baier and Bergstrand (2005) address the endogeneity problem using instrumental variables, Heckman's control-function techniques (Heckman, 1997), and panel data estimates. They find that the best method for estimating the effect of RTAs on bilateral trade flows is through differenced panel data using fixed effects, since a random effects estimation assumes zero correlation between unobservables and RTAs. On the other hand, the instrumental variables estimator applied to cross-section data are biased and underestimated. The main reason is that in cross-section it is difficult to identify variables that are correlated with the *RTA* dummy variable and uncorrelated with trade flows. Indeed, the most recent gravity model estimations tend to use panel data regression techniques, since cross-section and pooled regression models may be affected by the exclusion or mismeasurement of trading pair-specific variables (Baldwin, 2006).

In the following, in order to remove from the estimated effect size any possible bias due to the gold medal mistake, we introduce a *No-country effects* dummy equal to 1 if the original studies do not use country fixed effects to remove the cross-section bias. With regard to the typologies of data used, we introduce only two dummies—*Cross-section* and *Pooled*—in order to avoid collinearity problems. Finally, as far as the estimation methods are concerned, a *Random effects* dummy equal to 1 is associated with panel models estimated through this approach, and is expected to capture the upward bias implied by this methodology. Similarly, the dummy *Ols* is equal to 1 if estimates are obtained through simple OLS and 0 where estimates are obtained using other approaches (i.e. instrumental variables, Hausman–Taylor, etc.); in fact, OLS estimates may yield biased and inconsistent estimates due to omitted variables and selection bias. Trade between any pair of countries is likely to be influenced by certain unobserved individual effects and, if the unobserved effects are correlated with the explanatory variables, the coefficients of the latter may be higher because they incorporate these unobserved effects.

The *silver medal* mistake arises from the fact that different measures of bilateral trade flows are used. Although some studies focus on directional trade using data on bilateral imports or exports, the most frequently used measure is the average of bilateral trade flows. However, gravity models are usually estimated in log form: in this case, computing the wrong average trade (the arithmetic average corresponding to the log of the sums, rather than the geometric average corresponding to the sum of the logs) tends to overestimate the trade effects. Moreover, recall that the difference between the sum of the logs and the log of the sums increases for unbalanced trade flows (Baldwin,

© 2010 Blackwell Publishing Ltd

2006). Accordingly, we check whether employing the log of average bilateral trade flows rather than the average of the logs of the trade flows leads to significantly higher estimates of the RTA effects, by introducing a dummy *Log* equal to 1 if the dependent variable is computed as the log of average bilateral trade flows.

Another problem related to log specification is based on the existence of zero trade flows. Several methods have been proposed to tackle this issue: many empirical studies simply drop from the dataset the pairs with zero trade, and estimate the log-linear form using OLS. However, when the zero values are excluded, we face a selection problem. This has been handled in the literature in various ways: by applying the Heckman two-step procedure; by estimating the model using a Tobit estimator with $Trade_{ij} + 1$ as the dependent variable; or by employing a Poisson fixed effects estimator. A study by Egger (2005) compares four different estimators with regard to their suitability for cross-section gravity models. Egger recommends the Hausman–Taylor approach, which provides consistent parameter estimates, claiming OLS or the traditional REM to be biased. In the following, different methodological dummies (*Heckman*, *Tobit*, *Poisson*) deal with the selection bias and the presence of zero trade flows.

The *bronze medal* mistake refers to the (quite common) practice of deflating nominal trade values by the US aggregate price index. Given that there are global trends in inflation rates, this procedure probably creates biases via spurious correlations (Baldwin and Taglioni, 2006). Since time fixed effects are expected to offset this bias, in our analysis the *No-time effects* dummy is equal to 1 when time fixed effects are not included in the regression to control for global trends.

In addition to these medal mistakes, we would also draw attention to the widely cited model by Anderson and van Wincoop (2003), which derives a gravity equation that can be interpreted as a reduced form of a theoretically grounded trade model. Since this model is based on the assumption of constant trade costs, its application is only consistent with cross-section data analysis. Moreover, omission of the so-called "multilateral resistance" term may lead to inconsistent estimates. Studies that do not include the multilateral trade resistance term are characterized by an *Anderson–van Wincoop* dummy.

With regard to the dummies describing different features of the studies considered, we expect RTAs and their impact on trade to change over a period of time. Accordingly, we use four dummies—*Before 1970*, *1970s*, *1980s*, and *After 1990*—in order to collect studies using data related only to specific time periods. The *Dynamic* dummy refers to studies using dynamic techniques, though most of the papers in our sample rely on static panel gravity models. Finally, the *Agreement* dummy takes the value 1 if the original paper used a variable for each type of agreement while, as mentioned in the previous section, we handle extreme values in the sample by introducing the dummy *Outlier* (equal to 1 for outliers and 0 otherwise).

Regarding (possible) publication bias, we distinguish published from unpublished studies as well as papers primarily interested in estimating RTAs' impact on trade, from papers that include it as a mere control variable. Since we believe that published and very specific studies tend to include more accurate econometric analyses, we introduce a dummy *Unpublished* equal to 1 for not published papers, and a dummy *Control variable* equal to 1 for papers that insert the variable *RTA* simply as a control. It could be argued that the individuation of these biases is in the eye of the beholder, and our results could be seen as suspect. However, we follow Baldwin's (2006, p. 36) advice: "Please, suspect! That's what empirical researchers get paid for," and use these dummies without apology.

© 2010 Blackwell Publishing Ltd

Econometric Results

The use of a single observation for each study (sample of 85 estimates) raises the question of how to make the choice. Some authors identify a "preferred" estimate (Card and Krueger, 1995; Rose and Stanley, 2005) either randomly or using a more objective statistical procedure, such as the highest R^2 in the corresponding regression (Disdier and Head, 2004). Bijmolt and Pieters (2001) show that procedures using a single value for each study generate misleading results.

Indeed, if we look at the fixed and random effects estimates based on the study's minimum, median, and maximum estimates of γ, we obtain very different results (Table 1). In all cases, we reject the null hypothesis of estimate homogeneity, and the H^2 and I^2 statistics both confirm the results of the Q-test. All the confidence bounds are positive and strongly reject the null hypothesis of no effect. The lowest estimate (minimum estimates–random effects) implies an increase in trade of 12% ($e^{0.11} - 1 = 0.12$), while the highest estimate (maximum estimates–random effects) will be larger than 285% ($e^{1.35} - 1 = 2.85$). Given these results, and considering that we would otherwise lose valuable information, especially from studies that estimate gravity equations for multiple years, in the following we present the results obtained from the largest sample, including all available observations.

The mean RTAs effect of 1827 estimated coefficients in our database is 0.59, the median is 0.38 and although the majority of coefficients are clustered between 0 and 1 (only 312 estimates report negative effects), the estimated trade coefficients range from –9.01 to 15.41. We employ the Grubbs test in order to detect the existence of outliers (Disdier and Head, 2004) and find 38 extreme values that are dealt with by inserting the corresponding dummy variable in the MRA.

Table 2 presents the combined meta-estimates of γ together with tests for the lack of any effect and homogeneity of the data (Q, H^2, and I^2 statistics). All tests consistently reject the homogeneity hypothesis, and the heterogeneity among estimates leads to large differences among the fixed and random effects results. The null hypothesis is easily rejected, confirming the existence of a genuine impact of RTAs on bilateral trade. If we look at the sample of all published and unpublished estimated, the smaller fixed effects estimate indicates that RTAs raise trade by 10%, whereas the random effects estimate indicates an increase of up to 65%.

Table 2 also compares the results of the MA applied to the two subsets of published and unpublished research.[5] If published papers are deemed to have stronger methodological and theoretical foundations, the MA applied to the subset of published research should provide a more reliable estimate of the impact of RTAs. In this respect, if we limit our analysis to published papers, the results are significantly lower impacts, especially in the case of the fixed effects estimate. Lower values for the estimates for the published literature may be a positive result, suggesting that editors do a fairly good job at excluding the highest (and possibly less realistic) analyses. In contrast, the random effects estimates for the two subsets are much closer, as might be expected given that this type of estimation reduces the relative weighting given to more precise results, which are more likely in the case of published studies.[6]

Estimates of the impact of RTAs would seem to indicate a positive effect on trade, but the funnel graph diagram (Figure 1) clearly shows that the plot is overweighted on the right-hand side. The average of the top six points on the graph—that is, the estimates associated with the smallest standard errors—is equal to 0.04, implying a 4.1% increase in trade. Consequently, if research reporting were unbiased,

© 2010 Blackwell Publishing Ltd

© 2010 Blackwell Publishing Ltd

Table 1. Sensitivity to the Choice of Preferred Estimate

	Effects	*Pooled estimate*	*Lower bound of 95% CI*	*Upper bound of 95% CI*	*p-Value for H_0: no effect*	*Q-test (p-value)*	$H^2 = \frac{Q}{n-1}$	$I^2 = \frac{H^2-1}{H^2}$
Min.	Fixed	0.013	0.006	0.019	0.00	0.00	0.49	98%
	Random	0.113	0.049	0.178	0.00			
Median	Fixed	0.088	0.078	0.097	0.00	0.00	0.40	98%
	Random	0.531	0.455	0.608	0.00			
Max.	Fixed	0.414	0.400	0.427	0.00	0.00	0.99	99%
	Random	1.354	1.188	1.520	0.00			

Table 2. MA of Published and Unpublished Estimated RTAs Effect on Trade

Sample	*Effects*	*Pooled estimate*	*Lower bound of 95% CI*	*Upper bound of 95% CI*	*p-Value for H_0: no effect*	*Q-test (p-value)*	$H^2 = \frac{Q}{n-1}$	$I^2 = \frac{H^2-1}{H^2}$
1827 individual estimates	Fixed	0.100	0.097	0.101	0.00	0.000	0.48	98%
	Random	0.500	0.482	0.515	0.00			
731 published estimates	Fixed	0.055	0.053	0.058	0.00	0.000	0.56	98%
	Random	0.475	0.453	0.496	0.00			
1096 unpublished estimates	Fixed	0.218	0.214	0.222	0.00	0.000	0.38	97%
	Random	0.510	0.483	0.538	0.00			

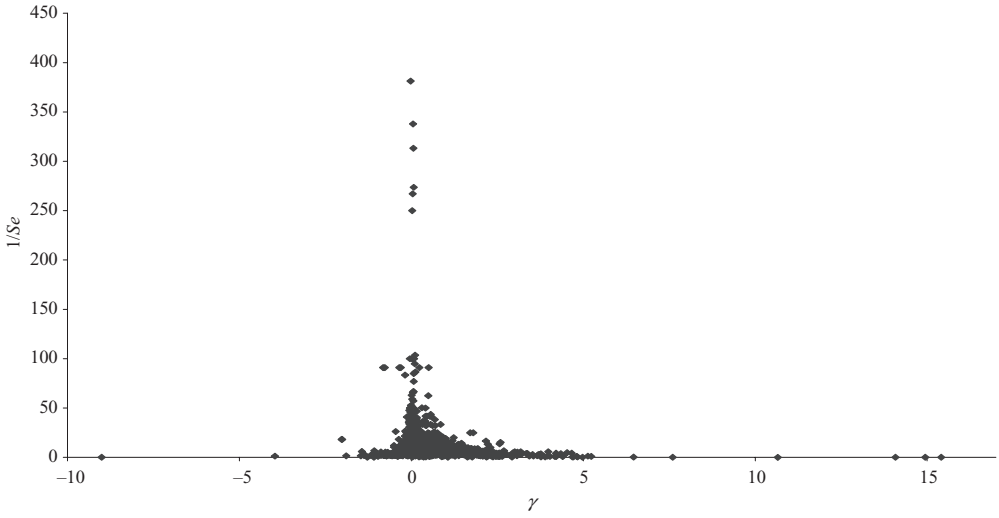

Figure 1. Funnel graph of 1827 individual estimates

estimates would vary randomly and symmetrically around the value 0.04, whereas the simple average for all 1827 estimates is 0.59, implying an 80% increase in trade.

Table 3 presents the result of the MRA tests and the standard errors adjusted for the 85 studies/clusters. Robust OLS estimations are given, and standard errors are recorded in parentheses. Model 1 is based on equation (5), dropping the insignificant variables one at a time, while Model 2 presents the estimated coefficients with the introduction of fixed effects for each type of agreement. If we compare the two models, the results are by and large robust. The estimate of β_0 confirms the apparent asymmetry of the funnel graph, since the reported effect is not independent of its standard error, whereas the β_1 estimate provides evidence of a significant general RTA effect on trade of around 40%.

The large negative coefficient associated with the *No-country effects* dummy highlights the downward bias in the studies that do not include fixed effects to characterize the trade flows in a particular country. The results are also negative for the *Cross-section* and *Pooled* variables: these classes of models, then, would seem to be affected by the exclusion or mismeasurement of trading pair-specific variables (Baldwin, 2006). Accordingly, our results support the claim made by Baier and Bergstrand (2005) that cross-section estimates are downward biased due to the endogeneity problem. On the other hand, we find a positive and significant coefficient for the *Random effects* and *Ols* dummies, and this provides an estimate of the upward bias due to the assumption of zero correlation between unobservables and RTAs. Finally, the omitted variables problems associated with the gold medal mistake can seriously affect the estimation of RTA trade impacts in both directions. From the coefficient values, though, it would appear that, overall, there is some evidence of a significant downward bias on the estimated impacts.

The confusion between the log of the average and the average of the logs tends to inflate the gravity estimates, leading to significantly higher estimates of the RTAs effect. This result confirms and provides a quantitative assessment of the silver medal mistake highlighted by Baldwin and Taglioni (2006).

© 2010 Blackwell Publishing Ltd

Table 3. MRA of RTA Effects

	Coefficient (robust with cluster standard errors)	
Variables	*Model 1*	*Model 2*
β_0: *Intercept*	3.34 (0.44)***	2.79 (0.48)***
β_1: $1/Se_i$	0.43 (0.14)***	0.39 (0.14)***
No-Country effects	–0.26 (0.10)***	–0.28 (0.11)***
Log	0.13 (0.06)**	0.15 (0.06)**
No-Time effects	0.15 (0.06)**	0.15 (0.06)**
Random effects	0.13 (0.07)*	0.16 (0.08)**
Cross-section	–0.22 (0.04)***	–0.21 (0.07)***
Pooled	–0.19 (0.04)***	–0.19 (0.05)***
Ols	0.21 (0.04)***	0.24 (0.05)***
Agreement	–0.11 (0.05)**	—
Control	–0.31 (0.07)***	–0.30 (0.07)***
Unpublished	0.10 (0.04)***	0.15 (0.05)***
Outliers	3.00 (0.27)***	2.52 (0.60)***
Before 1970	–0.35 (0.12)***	–0.29 (0.14)**
1970s	0.04 (0.22)	0.12 (0.24)
1980s	–0.21 (0.09)***	–0.16 (0.09)*
After 1990	–0.19 (0.05)***	–0.17 (0.06)***
Specific agreement effects	No	Yes
R-squared	0.25	0.34
Prob. > *F*	0.00	0.00
SE of regression	5.41	5.11

Notes: No. of obs. (no. of clusters) = 1827 (85); *** significant at 1%; ** significant at 5%; * significant at 10%.

The *No-time effects* dummy is expected to offset the bronze medal error implied by the incorrect deflation procedure. The positive sign associated with this variable shows that uncorrected studies also tend to overestimate the RTA impacts on trade.

As far as the variables related to each study's characteristics are concerned, we find a negative and highly significant coefficient for the *Agreement* dummy. Accordingly, studies focusing on specific RTAs tend to estimate much lower impacts on trade. In contrast, the estimated coefficient of the *Outlier* dummy is clearly positive since most extreme values are positive.

The positive coefficient found for the dummy *Unpublished* confirms the results in Table 2, while the dummy *Control* is strongly negative, hinting at the existence of a downward bias in studies that are not primarily interested in estimating the RTAs effect, and use this dummy as a control variable.

Finally, with the exception of *1970s*, we find significant and negative coefficients associated with the dummies for period ranges, so that the most recent studies seem to involve lower absolute values. This result is consistent with the often noted evolution from "shallow" to "deep" regional integration agreements, where the latter reduce trade costs through behind-the-border reforms.

In terms of individual RTAs, it emerges that 46 studies out of 85 estimate the impact of RTAs on trade introducing a different dummy for each trade agreement, yielding 1338 estimates. The largest number of observations refers to the European Union, one of the longest established and most studied cases of economic integration. Apparently,

© 2010 Blackwell Publishing Ltd

Table 4. MA of Estimates of Specific RTAs

Reciprocal trade agreements	*Effects*	*Pooled estimate*	*ΔTrade (%)*	*Q-test (p-value)*	*No. of estimates*	*Mean*	*Min.*	*Max.*
Association of Southeast Nations ASEAN (AFTA)	Fixed	0.67	95	0.00	41	0.81	–0.07	2.35
	Random	0.79	120					
Anglo-Irish Free Trade Area Agreement (AIFTAA)	**Fixed**	**0.07**	**7**	**0.18**	**10**	**0.06**	**0.00**	**0.10**
	Random	**0.07**	**7**					
Australia–New Zealand Closer Economic Relations (ANZCER)	Fixed	0.73	107	0.00	15	0.87	–0.16	3.98
	Random	0.88	142					
Baltic Free Trade Area (BFTA)	**Fixed**	**3.03**	**1972**	**0.04**	**24**	**2.96**	**2.37**	**3.77**
	Random	**3.06**	**2026**					
Central American Common Market (CACM)	Fixed	0.34	40	0.00	37	1.19	0.01	4.40
	Random	1.03	179					
Andean Community of Nations (CAN)	Fixed	1.10	200	0.00	13	1.34	0.12	2.22
	Random	1.23	242					
Caribbean Community (Caricom)	Fixed	0.29	34	0.00	37	2.02	–0.35	5.23
	Random	1.69	440					
Central Europe Free Trade Agreement (CEFTA)	Fixed	0.26	30	0.00	57	0.41	–0.51	1.52
	Random	0.40	49					
Commonwealth of Independent States Customs Union (CISCU)	Fixed	2.94	1795	0.02	6	2.66	1.98	3.37
	Random	2.82	1581					
Canadian–US Trade Agreement (CUSTA)	**Fixed**	**–0.34**	**–29**	**0.00**	**63**	**–0.23**	**–1.89**	**2.26**
	Random	**–0.25**	**–22**					
European Free Trade Association (EFTA)	Fixed	0.05	6	0.00	343	0.23	–1.38	2.17
	Random	0.24	27					
European Union (EU)	Fixed	0.05	6	0.00	524	0.52	–9.01	15.41
	Random	0.35	41					
Latin American Free Trade Agreement (LAFTA)	Fixed	1.14	213	0.00	5	0.98	0.30	2.57
	Random	0.98	168					
Latin American Integration Agreement (LAIA)	**Fixed**	**0.52**	**68**	**0.13**	**9**	**0.53**	**0.39**	**0.82**
	Random	**0.52**	**69**					
Southern Common Market (Mercosur)	Fixed	0.37	45	0.00	47	0.72	0.12	4.35
	Random	0.64	90					
North American Free Trade Agreement (NAFTA)	Fixed	0.80	123	0.00	90	0.90	–1.47	3.89
	Random	0.84	131					
US–Chile	Fixed	0.13	14	0.00	5	0.27	–0.30	1.42
	Random	0.27	31					
US–Israel	Fixed	0.80	122	0.00	12	0.82	–0.08	2.41
	Random	0.84	131					

the range between minimum and maximum estimates is very large for most agreements, demonstrating the large variety of estimates provided in the literature. Table 4 presents the results of the MA for the RTAs for which estimates are available.

The tests show that random effects estimates would be the most appropriate in most cases. Only four out of the 18 agreements do not show significant differences between fixed and random effects estimates (in bold in Table 4) and most of these cases are characterized by a fairly low number of observations. The largest effect is registered for the Baltic Free Trade Area where estimates suggest an increase in trade of around 2000%! Other agreements presenting exceedingly high estimates are the Commonwealth of Independent States Customs Union (CISCU)—1581%—and the Caribbean Community (Caricom)—400%. Looking at the most widely studied agreements—European Union (EU), European Free Trade Area (EFTA), and North American Free Trade Agreement (NAFTA)—the largest impact is for NAFTA (131%); the European

© 2010 Blackwell Publishing Ltd

Table 5. Probit Analysis

	Probit regression	
Independent variables	*Estimated coefficients (standard errors)*	*Marginal effects*
Before 1970	–0.92 (0.16)***	–0.35
1970s	–0.57 (0.16)***	–0.22
1980s	–0.82 (0.12)***	–0.32
After 1990	–0.25 (0.10)***	–0.10
No-Country effects	0.29 (0.12)**	0.11
Log	–0.45 (0.21)**	–0.16
No-Time effects	–0.27 (0.11)**	–0.10
Anderson–van Wincoop	–0.59 (0.14)***	–0.21
Random effects	0.53 (0.22)***	0.19
Pooled	0.54 (0.12)***	0.20
Cross-section	0.42 (0.12)***	0.16
Ols	–0.44 (0.10)***	–0.17
Heckman	–0.44 (0.26)*	–0.18
Tobit	–0.83 (0.17)***	–0.32
Poisson	–0.68 (0.19)***	–0.27
Dynamic	–0.54 (0.17)***	–0.21
Agreement	–0.50 (0.09)***	–0.19
Unpublished	–0.20 (0.08)***	–0.08
Control	–0.45 (0.08)***	–0.17
Intercept	2.51 (0.27)***	–0.35
Wald χ^2 (19)	340	
(p-Value)	0.00	
Pseudo	0.14	

Notes: No. of obs: 1827; *** significant at 1%; ** significant at 5%; * significant at 10%.

agreements register much lower, but possibly more realistic, values: 27% for EFTA and 41% for the EU. It is also worth noting that customs unions, such as EU, Caricom, Southern Common Market, Central American Common Market and CISCU, do not seem to consistently outperform the free-trade areas in terms-of-trade impact.

Probit Significance Equation

In our dataset of 1827 effect sizes, 1134 are significantly different from zero at the level of 5%, and 1049 of these estimates are positive. This is the sample used in the probit estimates (equation (6)) which include the set of variables already presented. Results in terms of marginal effects are presented in Table 5. Comparing these estimates with those in Table 3 we can identify three groups of variables: significant variables in both cases, with the same sign; significant variables in both cases, with opposite signs; significant variables in the probit regression that were dropped from the MRA.

In the first group, we find the dummies for different decades, the analysis of specific RTAs, the use of an REM in panel estimations, and the papers that do not focus on RTAs. In these cases, then, the probit estimates confirm the evidence provided by the MRA. First, the assessments of older agreements (or those in the first stages of implementation) are less likely to detect a positive impact on trade: using data after 1990,

© 2010 Blackwell Publishing Ltd

for instance, reduces the probability of positive and significant results by 10% while the coefficient before 1970 is 35%. By the same token, using the data on specific agreements and the *RTA* dummy as a simple control substantially reduces the probability of estimating a positive impact on trade by 19% and 17%, respectively, which is as expected given that the estimates provided by these studies are generally lower. On the other hand, panel estimates based on random effects raise the probability of finding a positive and significant effect.

In the second group, we find some dummies related to the "medal errors" (*No-country effects*, *Log*, *Ols*, and *No-time effects*), data used (*Cross-section* and *Pooled*), and publication impact (*Unpublished*). In almost all cases, the probit estimates show that possible errors and the biases previously mentioned tend to decrease the probability of a significant and positive impact on trade notwithstanding the overestimation highlighted in Table 3. On the other hand, those studies that do not include fixed effects to characterize the trade flows involving a particular country are more likely to generate fallacious positive estimates while the downward bias indicated by the MRA is mostly due to negative estimates.

In the third group, we find some methodological dummies that do not have a significant impact if we use the larger sample. As might be expected, failure to take account of the multilateral trade resistance term (*Anderson–van Wincoop*) decreases the probability of a positive and significant estimate; while use of more sophisticated estimation methods (*Dynamic*) dealing with the selection bias and the presence of zero trade flows (*Heckman*, *Tobit*, *Poisson*), decreases the probability of "false positive" results.

4. Conclusion

RTAs have been widely studied, and the interest on this type of trade liberalization is likely to increase in the immediate future due to the crises in the multilateral liberalization process. One way to carry out a comparative study of empirical results is to simply tabulate authors, countries, methodology, and results. However, for policy analysis and to achieve a better understanding of the consequences of RTAs, it is useful to complement broad qualitative conclusions with a more precise quantitative research synthesis. This is the aim of the present paper with respect to one core issue: assessment of the impact of these agreements on member countries' bilateral trade flows using gravity models. In particular, we try to overcome the main limitations of qualitative reviews by summarizing the whole body of work through MRA.

The estimated effects of RTAs vary widely, from study to study and sometimes even within the same study. From a methodological point of view, this suggests the possibility of retaining all the available observations in most of the statistical analysis, but considering estimates from the same study as possibly correlated observations. Accordingly, by means of meta-analysis techniques, we statistically summarized 1827 estimates collected from a set of 85 studies. Our results have dual value-added.

First, all estimates combined imply a substantial impact on trade, since the random effects estimate demonstrates an increase of 65%. The more modest fixed effects estimate (10%) cannot be relied upon because its basis is undermined by obvious heterogeneity in this literature. After filtering out publication impact and other biases, the MA confirms a robust, positive RTAs effect, equivalent to an increase in trade of around 40%. The estimates tend to get larger for more recent years, which could be a consequence of the evolution from "shallow" to "deep" trade agreements. Looking at the effects for type of trade agreement, we find evidence of a differen-

© 2010 Blackwell Publishing Ltd

tiated impact on trade and, indeed, in many cases the MA estimate largely exceeds the estimate for all the agreements combined. Overall, there is evidence that *ex post* empirical estimates of an influence of RTAs on trade flows are positive and nontrivial.

Second, the MRA provides a range of additional results which help to explain the large variation in reported estimates. Although there is still a large unexplained variation in the meta-regression models, our results shed some light on the role played by certain research characteristics in explaining this variation. In this respect, MA statistical techniques are more than mere weighted averages of all the point estimates, since they provide a quantitative assessment of the consequences of publication selection, and possibly questionable methodological choices. As far as the latter are concerned, there appears to be evidence of a significant downward bias due to omitted variables problems (gold medal mistake), while data measurement (silver medal mistake) and specification problems (bronze medal mistake) are less likely to produce (statistically speaking) "good results," and estimates tend to be biased in the opposite direction.

Regarding publication selection, in considering all collected point estimates we do not dispense with any (possibly) valuable information. However, we do have a view about the quality of some of these estimates and this impinges upon our interpretation of the results. In this respect, estimates obtained from gravity models using the *RTA* dummy as (some other) control variable are largely downward biased and are much less likely to produce significant results. More importantly, our results fly in the face of the general consensus among meta-analysts that referees and editors are predisposed to treating "large and significant" results more favorably. In the literature that we reviewed, there is strong statistical evidence of a nonpublication bias which favors the reporting of positive trade effects, while the publication process leads to lower, and probably more realistic, estimates.

References

Anderson, J. E. and E. van Wincoop, "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review* 93 (2003):170–92.

Ashenfelter, O., C. Harmon, and H. Oosterbeek, "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias," *Labour Economics* 6 (1999): 453–70.

Baier, S. L. and J. H. Bergstrand, "Do Free Trade Agreements Actually Increase Members' International Trade?" Federal Reserve Bank of Atlanta, working paper 2005-3 (2005).

Baldwin, R., *Towards an Integrated Europe*, London: Centre for Economic Policy Research (1994).

———, "The Euro's Trade Effects," European Central Bank, working paper 594 (2006).

Baldwin, R. and D. Taglioni, "Gravity for Dummies and Dummies for Gravity Equations," NBER working paper 12516 (2006).

Bhagwati, J., P. Krishna, and A. Panagariya (eds), *Trading Blocs*, Cambridge, MA: MIT Press (1999).

Bijmolt, T. H. A. and R. G. M. Pieters, "Meta-Analysis in Marketing when Studies Contain Multiple Measurements," *Marketing Letters* 12 (2001):157–69.

Card, D. and A. B. Krueger, "Time-Series Minimum-Wage Studies: A Meta-Analysis," *American Economic Review* 85 (1995):238–43.

Clarete, R., C. Edmonds, and J. S. Wallack, "Asian Regionalism and its Effects on Trade in the 1980s and 1990s," *Journal of Asian Economics* 14 (2003):91–129.

De Dominicis, L., H. de Groot, and R. Florax, "Growth and Inequality: A Meta-Analysis," Tinbergen Institute, discussion paper TI 2006-064/3 (2006).

© 2010 Blackwell Publishing Ltd

Disdier, A. C. and K. Head, "The Puzzling Persistence of the Distance Effect on Bilateral Trade," Centro Studi Luca D'Agliano, Development Studies, working paper 186 (2004).
Egger, P., "Alternative Techniques for Estimation of Cross-Section Gravity Models," *Review of International Economics* 13 (2005):881–91.
Eichengreen, B. and D. Irwin, "The Role of History in Bilateral Trade Flows," NBER working paper 5565 (1996).
Feenstra, R. C., "Integration of Trade and Disintegration of Production in the Global Economy," *Journal of Economic Perspectives* 12 (1998):31–50.
Florax, R., "Accounting for Dependence among Study Results in Metaanalysis: Methodology and Applications to the Valuation and Use of Natural Resources," Research Memorandum 2002.5, Faculty of Economics and Business Administration, Free University, Amsterdam (2002).
Ghosh, S. and S. Yamarik, "Are Regional Trading Arrangements Trade Creating? An Application of Extreme Bounds Analysis," *Journal of International Economics* 63 (2004):369–95.
Glass, G. V., B. McGaw, and M. Lee Smith, *Meta-Analysis in Social Research*, Beverly Hills, CA: Sage (1981).
Heckman, J. J., "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32 (1997):441–62.
Higgins, J. P. T. and S. G. Thompson, "Quantifying Heterogeneity in a Meta-Analysis," *Statistics in Medicine* 21 (2002):1539–58.
Jarrell, S. B. and T. D. Stanley, "A Meta-Analysis of the Union Wage Gap," *Industrial and Labour Relations Review* 44 (1990):54–67.
Jeppesen, T., J. A. List, and H. Folmer, "Environmental Regulations and New Plant Location Decisions: Evidence from a Meta-Analysis," *Journal of Regional Science* 42 (2002):19–49.
Lawrence, R. Z., *Regionalism, Multilateralism, and Deeper Integration*, Washington, DC: Brookings Institution (1996).
Meade, J., *The Theory of Customs Unions*, Amsterdam: North-Holland (1955).
Rose, A. K. and T. D. Stanley, "Meta-Analysis of the Effect of Common Currencies on International Trade," *Journal of Economic Surveys* 19 (2005):347–65.
Stanley, T. D., "Wheat from Chaff: Meta-Analysis as Quantitative Literature Review," *Journal of Economic Perspectives* 15 (2001):131–50.
———, "Beyond Publication Bias," *Journal of Economic Surveys* 19 (2005):309–45.
Stanley, T. D. and S. B. Jarrell, "Meta-Regression Analysis: A Quantitative Method of Literature Surveys," *Journal of Economic Surveys* 19 (2005):299–308.
Sutton, A. J., K. R. Abrams, D. R. Jones, T. A. Sheldon, and F. Song, *Methods for Meta-Analysis in Medical Research*, Chichester: John Wiley (2000).
Viner, J., *The Customs Union Issue*. New York: Carnegie Endowment for International Peace (1950).
World Bank, "Regional Trade Agreements: Effects on Trade," ch. 3 in *Global Economic Perspectives*, Washington, DC: World Bank (2005).

Notes

1. An anonymous referee pointed out that if gravity models tend to be biased in a particular direction due to a common misspecification, our meta-analysis estimates are going to include the average of this systematic bias.
2. A complete list of the full sample of papers and detailed information on estimates are available from the authors.
3. A moment-based estimate of $\hat{\tau}^2$ may be obtained by (8) equating the observed value of Q with its expectation

$$\mathrm{E}[Q] = \hat{\tau}^2\left(\sum_{i=1}^{n} w_i - \left(\sum_{i=1}^{n} w_i^2 \Big/ \sum_{i=1}^{n} w_i\right)\right) - n + 1.$$

4. According to Stanley (2005, p. 310): "Publication bias, or the 'file drawer problem', has long been a major concern to meta-analysts. In its more benign form, it is the result of selection for

© 2010 Blackwell Publishing Ltd

statistical significance. Researchers, reviewers, and editors are predisposed to treat 'statistically significant' results more favorably; hence, they are more likely to be published. Studies that find relatively small and 'insignificant' effects tend to remain in the 'file drawer'."

5. This comparison was suggested by an anonymous referee.

6. A within-study MA of RTA effects on trade for each of the 85 studies shows that the null hypothesis of no effect is easily rejected at any standard significance level for most of the studies in both subsets. Results are available from the authors upon request.

© 2010 Blackwell Publishing Ltd