Do We Really Know the WTO Cures Cancer? False Positives and the Effects of Institutions[*]

Stephen Chaudoin
University of Illinois at Urbana-Champaign
Jude Hays
University of Pittsburgh
Raymond Hicks
Princeton University

December 16, 2014

[*] We appreciate helpful advice from Marc Busch, Jake Bowers, William Clark, Mark Fredrickson, Kristian Skrede Gleditsch, Sarah Hummel, In Song Kim, Moritz Marbach, Michael Miller, and Dustin Tingley. We also appreciate comments from audiences at the International Political Economy Society, American Political Science Association, and International Studies Association conferences.

Abstract

Assessing the effect of institutions on state behavior is a difficult task since unobservable factors affect institutional membership and compliance with the institution's rules. These unobservables potentially bias estimates in favor of finding a potentially false positive effect of institutions. We use a replication experiment of 94 specifications from 16 different studies to show the severity of this problem. Using a variety of existing approaches, we show that membership in the GATT/WTO institution has a significant effect on a surprisingly high number of dependent variables (34%), variables which have little to no theoretical relationship to the multilateral trade regime. Monte Carlo simulations confirm that the problem is severe even in controlled environments. We apply two types of sensitivity analysis and give guidance for the conditions under which each sensitivity approach can guard against false positives.

Does ratifying or joining an international institution have a causal effect on a country's policies? This vein of research encompasses critical questions such as whether human rights treaties improve human rights, whether free trade agreements increase trade, or whether alliances change conflict behavior. Generally, scholars ask whether member states change their policies to be in line with an institution's rules, i.e. compliance.

Assessing the relationship between ratification and compliance is difficult because the same factors that drive compliance also drive the initial decision to join an institution. Often these factors are unobservable, meaning that they either are not easily measured or known to the researcher. This problem, which is often called "selection on unobservables," most likely biases empirical findings regarding the effects of institutions in a positive direction, because countries who are most likely to comply *ex ante* are also the most likely to ratify (Downs, Rocke and Barsoom, 1996). Even if ratification has no causal effect on compliance, selection on unobservables can result in "false positives," where estimates incorrectly suggest a positive effect of ratification on compliance. When we observe a positive relationship between ratification and compliance, we are left wondering whether this finding reflects a true causal relationship, or if is only an artifact of selection on unobservables.

Extant IR research uses a veritable smorgasbord of empirical models designed to address this problem. We ask: do these fixes work? In other words, when we employ these empirical estimation approaches, can we be confident that a positive finding represents a causal relationship between membership and compliance, as opposed to a false positive?

We present evidence from an extensive replication exercise that the answer is no. Specifically, we start with a set of existing studies which analyze dependent variables which are *not* closely linked theoretically to international trade, e.g. a country's torture rate or whether it has a legislature. Using identical models to the authors' original specifications, we add a variable coding the country's membership in the World Trade Organization (WTO) to assess whether WTO membership had a statistically significant effect on those dependent variables, despite there being virtually no theoretical relationship between WTO membership and those dependent variables. We find a disconcertingly high rate of significant results. The WTO has a statistically significant relationship approximately 34% of the time, which is over three times as high as the rate implied by conventional levels of statistical significance. We also show how the most commonly used estimation approaches do not reduce these false positive rates, and in some instances, make the problem worse by creating new false positives where there were none before.

To be sure, it is impossible to know whether a particular result represents a false positive or a true relationship. To address this, we make our replication exercise even more conservative. We show how our results obtain using a treaty that has an even more tenuous theoretical link with the dependent variables we consider: the Convention on Trade in Endanged Species (CITES). It is very unlikely that CITES, a convention which institutes licensing requirements for a small amount of trade in endangered plants and animals, has a causal relationship with the dependent variables in our replication exercise, none of which describe environmental outcomes. This gives convincing evidence that our WTO findings are not merely the result of true relationships which researchers do not yet understand. Together, these results should give researchers pause regarding the degree to which extant approaches address the problem of selection on unobservables.

While the WTO and CITES replication exercise diagnoses the potential severity of false positives, the second section explains the problem in a controlled environment. We present a generic data generating process (DGP) that highlights the key features of the selection on unobservables problem. Unobservables can take many different types. Some are country-specific and time-invariant. Others are time-varying, but common across countries. Still others are country-specific and time-varying. Each type is theoretically plausible and supported by arguments in existing literature. Yet each also has different implications for the conditions under which existing fixes are susceptible to generating false positives.

We conduct Monte Carlo simulation analysis in which we vary the type and strength of unobservables and show which types of approaches work best under different conditions. Because we set up the the DGP to explicitly reflect selection on unobservables, we can be confident that this is the problem and the simulations do confirm our results from the replications. Even when the simulated data generating process is simpler than that of the real world and known to the researcher, false positives rates are high. The simulations also confirm the second, subtler result from the replication exercise. We demonstrate a "law of second best," where addressing one type of selection on unobservables can exacerbate the problems caused by the presence of other types. This helps explain why, in the replication exercise, different fixes, and their combinations, both created and removed false positives.

This paper is not all doom and gloom. In the final section, we show how sensitivity analysis is a powerful tool for assessing the likelihood that a positive result is a false positive. We show two types of sensitivity analysis. The first, comes from Altonji, Elder and Taber (2005) and leverages selection on observables as a guide for selection on unobservables. The intuition behind this approach is to ask “How severe would selection on unobservables need to be, relative to selection on observables, to account for the estimated effect of ratification on compliance?” The second approach, from Imbens (2003), leverages the power of observables to explain compliance to benchmark the likelihood of a false positive. We describe each approach mathematically and demonstrate each with examples from our replication studies.

Most importantly we give guidance on when researchers should apply each approach. When the researcher has stronger theoretical, prior knowledge about the ratification process, the first approach is more useful. When the researcher instead knows more about the compliance process, the second approach is more useful.

WTO membership and compliance are the examples of particular treatments and outcomes that we analyzed here, yet none of our arguments are idiosyncratic to that context. The problem of selection on unobservables is challenging for a wide array of political science applications. In comparative politics, researchers ask whether political and financial institutions, like democracy or central bank independence, affect outcomes like growth and inflation. It is possible that unobservables, e.g. a country’s overall stability or inflation aversion, affect both domestic institutions and outcomes. In American politics, researchers ask whether electoral rules affect turnout or whether higher court rulings affect lower court compliance. It is possible that unobservables, e.g. civic engagement or the strength of a legal argument, affect both rules and rulings, turnout and compliance. These are analogous hurdles to those faced by researchers assessing the effects of international institutions.

Furthermore, we have described our arguments in terms of false positives, because we have theoretical expectations that selection on unobservables biases estimates in a positive direction in this particular context. But our arguments apply generally to the bias in estimated effects that results from selection on unobservables, which may be positive or negative in other contexts. The characterization of the selection on unobservables problem, the sensitivity tests described, and the advice given here should be useful to scholars across subfields and applications.

The Problem of False Positives

A large body of IR research theorizes about whether and how international institutions cause sovereign nations to change their behavior. To test these theories empirically, researchers model the relationship between an explanatory variable that describes a country’s status vis-a-vis a particular institution and a dependent variable that describes some aspect of the country’s behavior or its policies. Most often, the explanatory variable measures whether a country has ratified or joined a particular treaty or organization. For the dependent variable, we are often interested in whether a country has adopted policies that are consistent with that institution's rules, often called compliance.

Examples abound in all areas of international relations research. In IPE, researchers ask whether the institutions governing international trade and finance affect government policies or economic outcomes. For example, Simmons (2000); Simmons and Hopkins (2005); Von Stein (2005) debated whether accepting the IMF's Article VIII commitments decreases a government's probability of implementing current account restrictions. A large body of work asks whether bilateral investment treaties affect investment. In human rights, much research asks whether membership in the Convention Against Torture and other legal instruments of international law affects a country's human rights policies. In conflict and security studies, many studies ask whether alliance membership affects a country's conflict behavior.

The empirical tests employed by researchers generally resemble the system described in Equation 1. r_{it} is a binary variable that equals one if country i has ratified a particular treaty in or before year t. c_{it} is a binary variable that equals one if country i's policies are compliant with the treaty's rules in year t. For simplicity, we will speak of countries as having ratified or not ratified a treaty, and their policies as either being in compliance with that treaty's rules or not.[1] The vector X_{it} contains the observable characteristics of a country which potentially affect compliance and ratification. u^r_{it} and u^c_{it} are unobservables that affect ratification and compliance respectively.[2]

$$\begin{aligned} r_{it} &= f(X_{it}B + u^r_{it}) && \text{(Ratification Equation)} \\ c_{it} &= f(X_{it}\beta + \alpha r_{it} + u^c_{it}) && \text{(Compliance Equation)} \end{aligned} \tag{1}$$

Researchers generally are interested in estimating α, the effect of ratification on compliance. In estimating α, researchers face a familiar problem: the unobservables that affect ratification are correlated with the unobservables that affect compliance, which biases estimates of α. In the context of treaty ratification and compliance, we usually think this correlation is positive, which

[1] Compliance need not be binary. In later sections, we consider both continuous and binary measurements of compliance.

[2] Of course, the particular functions used, $f()$, vary across estimation procedures. Some estimators do not use the linear and additive form described here. Our point is to demonstrate the basic moving parts of the problem.

biases estimates upwards. As a consequence, even when we find positive estimates of α, as are often predicted by theory, we should be suspicious about whether these are "true positive" findings or if they are "false positives," estimates which are artifacts resulting from correlation among unobservables.[3]

Possible False Positives

How likely are existing estimation approaches to generate false positive estimates of α, the effect of the institution on compliance? We find that false positives are very likely to be a problem. To support this claim, we use existing estimation approaches and see whether a particular treaty has significant effects on country level characteristics, despite there being little to no theoretical relationship between that treaty and those characteristics. The explanatory variable we use measures whether a country is a member of the GATT/WTO. The country level characteristics (dependent variables) that we analyze are quantities which are unlikely to be influenced by the multilateral trade regime, e.g. instances of torture, whether a country has a legislature, or literacy rates.

In the parlance of medical trials, this is like a placebo test. We take a set of patients, each of whom has a different disease (high torture, low literacy). We give each of them a placebo drug (WTO membership). And then we assess whether existing approaches would tell us that the placebo drug has an effect on the disease. By design, where we find statistically significant effects, we should be suspicious that they are false positives as opposed to true relationships between treatment and outcome. In the final part of this section, we analyze the Convention International Trade in Endangered Species, instead of the GATT/WTO regime. We do this as an even more conservative placebo test, since the theoretical link between CITES and the dependent variables analyzed here is virtually non-existent.

To be precise about language, from here forwards, "false positive" refers to a statistically significant relationship between the WTO/CITES and the outcome variable, not the sign of the coefficient. While our theoretical knowledge makes us suspect that the direction of bias resulting from selection on unobservables is positive in many situations, we focus here on the likelihood of finding any statistically significant relationship between WTO/CITES and outcomes, regardless of its direction.

[3] See Simmons (2000); Simmons and Hopkins (2005); Von Stein (2005). For a more recent treatment, see Lupu (2013).

Population of Studies

We began by gathering the population of studies published in *APSR, AJPS,* and *IO* from 2005-2013 that used a country-year unit of observation.[4] For each study, we identified the dependent variable, the set of explanatory variables, and the estimation procedure used to produce the published results. To standardize notation as we discuss these studies, let y_{it} denote the dependent variable of the study and let X_{it} denote the collection of explanatory variables. We then excluded studies which analyzed a dependent variable with a strong or potentially-strong theoretical link between WTO membership and that dependent variable.[5] Our explanatory variable, WTO_{it}, is a dummy variable that equals one if that country was a member of the GATT/WTO during that year and zero otherwise.

In all, we used 16 studies. For each study, we gathered the authors' replication data and replicated their analyses. Since there were multiple regressions/estimations in all the studies, this yielded a total of 94 replications.

Baseline Replications

For the baseline set of replications, we used authors' exact original specifications. The only change we made was to add the WTO_{it} variable as an additional explanator.

For each replication, we gathered the p-value associated with the coefficient on the WTO variable.[6] Figure 1 orders these p-values along the horizontal axis from least to greatest. The vertical axis shows the p-value for that particular replication. The horizontal red line marks the 0.10 level. The vertical red line marks the 32nd replication, which is the replication with the greatest p-value that still falls below the 0.10 threshold.

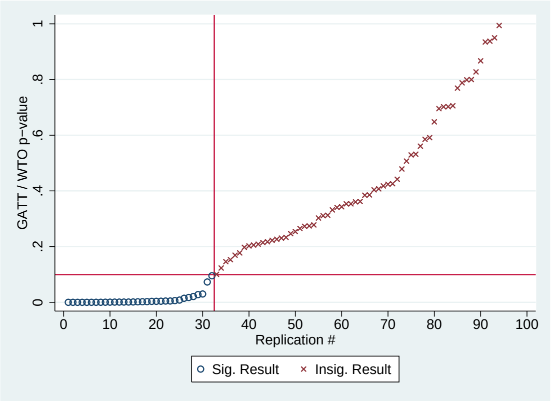
The two red lines divide the figure into four quadrants. Red X's in the top right correspond to "true negatives." These are studies where we would not expect to find any statistically significant effect for the WTO, and indeed do not. Blue O's in the bottom left correspond to "false positives,"

[4] We had to limit ourselves to studies where the authors provided replication materials online or upon request. We supplemented this set by also using some studies from *JOP, ISQ,* and *JCR* which used country-year units of observation and which also devoted significant attention to the problem of selection on unobservables. Including these additional studies should make our tests more conservative, because presumably, these studies are less susceptible to the problem of false positives. A full list of the studies is available in the appendix.

[5] We were conservative. Practically speaking, we excluded all trade-related dependent variables, e.g. trade, tariffs, etc.

[6] We calculated each p-value in the same way that the authors did, e.g. robust or clustered standard errors. We are interested in the likelihood that selection effects cause incorrect inferences, as opposed to the possibility that incorrect statistical calculations cause incorrect inferences. For work on the latter subject, see Bertrand, Duflo and Mullainathan (2004).

Figure 1: P-values for Effect of WTO on Irrelevant DV's



studies where the WTO has a statistically significant effect on the dependent variable.

The most important feature of the figure is that the overall false positive rate is much higher than we would expect. 32 replications have p-values less than 0.10, a false positive rate of approximately 34%. If using the conventional 0.10 critical level, we would expect to observe, by chance, approximately 9-10 significant results. We found over three times that number. The false positive results are also far from "barely significant." 30 of the replications have p-values less than 0.05. 25 of the replications have p-values less than 0.01.

The false positives are also not concentrated in just a few studies or estimation approaches. Of the 16 studies we replicated, almost half (7) had at least one replication in which the WTO variable was statistically significant. Of the 34 different dependent variables analyzed in the 16 studies, the WTO variable was statistically significant in at least one replication for 16 of the dependent variables. Some dependent variables were continuous while others were limited dependent variables. Of the 33 continuous dependent variable replications, the WTO variable was significant in 17 of them. Of the 61 limited dependent variable models, the WTO variable was significant in 15 of them.

Replications with Existing Fixes

Extant work uses a variety of approaches to address selection on unobservables. Some are based on panel data techniques used for unobserved heterogeneity and trending, like unit or year fixed effects, time trends or splines. Others have advocated matching techniques, based on the intuition that matching facilitates comparison of treated and control units which are similar to one another in terms of their observable characteristics.

For the second set of replications, we incorporated each of these different approaches. Some of the studies we replicated used these approaches in their published specifications, while others did not. Country fixed effects were the most commonly applied strategy for dealing with unobserved country-specific variation, used in 26 of the 94 replications. 72 of 94 used some sort of time-based fix, like splines, year trends or year fixed effects. 20 of the 94 used some combination of country fixed effects and time trends.

To assess the effect of these approaches on false positive rates, we began by stripping them out of all the replication specifications. We call these the "reduced" replications. They are identical to the authors' original specifications in every way except (a) we added the WTO_{it} variable and (b) we did not include any fixed effects, splines, etc.

We then applied each of these fixes one-by-one (and in combinations) to *all* replications. We can assess how the false positive rate changes as we apply certain types of fixes. Table 1 describes the number of false positives across these specifications. Column 1 provides the baseline results described above for comparison. Column 2 describes the reduced replication results. Column 3 adds country fixed effects to every replication (if they weren't already included) and removes any other fixes. Column 4 adds a country-specific linear time trend to any model that didn't already include some fix for time trends or period specific shocks. If the original model included a fix (time trend, year fixed effects, or splines) we left it in as specified by the author. For this column, we also removed any country fixed effects.

The final column of Table 1 describes the false positive rates from replications using a standard matching technique.[7] Matching techniques, in which the sample is pre-processed or pruned, are often used. In applied research, the most common justification for using this technique is to address non-random selection or endogeneity.

We use one of the most common matching techniques, propensity score matching.[8] Briefly, propensity score matching uses a set of observables to estimate the probability of a unit receiving treatment (GATT/WTO membership). Treated and untreated observations with similar propensity scores are matched together, and then the dependent variable is compared across the matched, treated and untreated observations to obtain an estimate of the effect of GATT/WTO membership.

Here, we used each of the covariates in the study to construct a propensity score, matched on that propensity score, and then calculated the average treatment effect of the treated observations. For choosing which variables to include in the propensity score matching procedure, we followed

[7] The p-values are computed using the post-processed sample size.

[8] Rosenbaum and Rubin (1983).

the advice of Ho et al. (2007): “All variables in X_i that would have been included in a parametric model without preprocessing should be included in the matching procedure” (216).[9] Each treated observation is matched with one other observation. The average treatment treatment effect on the treated is a weighted comparison of the mean of the dependent variable across treated and control units. When there are more treated units, control units which are matched with more treated units receive higher weights than control units which are not matched with many treated units. If there are more control units, again each treatment unit will receive a match, but control units might be matched more than once and some control units might not be matched.[10]

To be sure, there is much methodological debate and innovation over what variables to match on, how to match observations (propensity score, distance metrics, coarsening, etc.), and how to assess balance on observables after matching. Since our goal is not to weigh in on these debates, we would note that matching procedures are valuable techniques for achieving and assessing balance on observables. Yet even when achieving balance on observables, it is still possible for inferences to be biased because of imbalance on unobservables. For example, in the simulations in later sections, we can achieve very good balance on unobservables with a variety of matching procedures and our estimated ratification effects will still be biased as a result of selection on unobservables. For this reason, we focus on a standard, commonly used approach, rather than on variation in false positive rates across matching procedures.

Table 1: False Positive Rates for Replications, GATT/WTO Variable

	Specification				
	Orig.	Reduced	Country FE	Splines/Country Trend	Matching
False Pos. Rate	34%	44%	34%	34%	31%
No. Replications	94	94	91	94	90
No. Studies	16	16	16	16	16
Country Fixed Effects			26/94		
Time Trend?			72/94		
Limited Dep. Variable			62/94		

There are two important results from Table 1. First, the high rate of false positives is surprisingly persistent. The false positive rate rises from 34% to 44% when we remove the authors’ fixes. However, adding country fixed effects or country trends/splines only reduces the rate to 34% for both. The matching approach fares similarly, with a false positive rate of 31%.

[9] Others have advocated matching on observables which predict treatment. It’s worth noting that many of the replication studies’ observables included “standard” controls, like GDP or democracy, which are strong predictors of GATT/WTO membership as well.

[10] We used *psmatch2* in Stata, Leuven and Siansei (2003).

The second result from Table 1 is that fixes fix some problems, but also create new ones. Using particular fixes, many of the false positives in the baseline replications are removed. Some replications which previously generated significant results now generate insignificant results. However, the fixes create new false positives where there were none before.

Figure 2 shows the p-values for the country fixed effects replications. For this figure, we kept the ordering of the studies the same as in Figure 1 and we retained the same vertical and horizontal red lines. For Figure 2, red X's still denote insignificant p-values, greater than 0.10, and blue O's still denote significant p-values, less than 0.10.

Figure 2 shows how country fixed effects ameliorate the false positives problems in some ways and exacerbate it in others. There are 9 red X's in the upper left quadrant of the figure, which denote the 9 replications in which the GATT/WTO variable was significant without country fixed effects, but is no longer significant with country fixed effects. This is encouraging- these are replications where the GATT/WTO variable becomes insignificant with a commonly applied fix. However, there are also 8 blue O's in the bottom right quadrant. These are new false positives: studies for which the WTO variable was insignificant without country fixed effects, but is now significant with country fixed effects.

Figure 3 shows the same results using the matching replications. There are 14 red X's in the top left quadrant- studies where the GATT/WTO variable was significant, but is insignificant when we use matching. However, there are 12 blue O's in the bottom right quadrant- new false positives that arise from the matching approach.

The false positives from the matching replications also were not simply caused by a failure to achieve balance on observables. The degree to which the matching procedure achieved balance on observables varied across replications. However, better balance was not associated with a decreased false positive rate. The mean percent reduction in bias, averaged across each of the observables used in the replication, was very similar for replications which did and did not result in a positive result. A simple regression of the probability of a false positive on the percent reduction in bias shows virtually no association between the two.[11] And to reiterate, in the later simulations sections, we can show that high false positive rates due to selection on unobservables can obtain even when achieving a very high level of balance on observables.

[11] The logit coefficient on the percent reduction in bias is 0.001 with a p-value of 0.941.

Figure 2: P-values for Effect of GATT/WTO on Irrelevant DV's, Fixed Effects

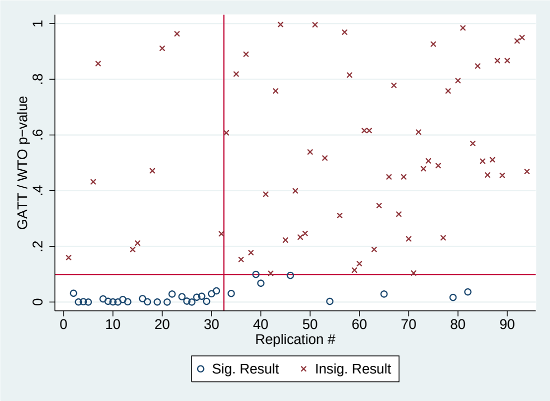
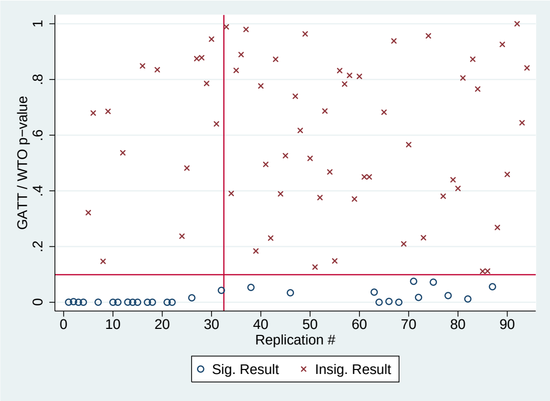


Figure 3: P-values for Effect of GATT/WTO on Irrelevent DV's, Matching



Combining Fixes

Table 2 shows that *combinations* of fixes also fail to lower the false positive rate. Column 1 strips out any existing time-based fixes and includes a country-specific linear trend in each replication. Column 2 repeats this and also adds country fixed effects. Column 3 is identical to Column 1, only it uses year fixed effects instead of country specific linear trends. Column 4 uses country and year fixed effects.

The false positive rate is lowest when using country specific linear trends in isolation, as in Column 1. Yet, even this is almost twice the rate afforded by conventional levels of statistical significance. Adding country and/or year fixed effects raises the false positive rates back to rates closer to Table 1.

Table 2: False Positive Rates for Replications with Multiple "Fixes," GATT/WTO Variable

	Specification			
	Cty. Trends	Cty. Trends + Cty. FE	Year FE	Cty. and Year FE
False Pos. Rate	17%	25%	35%	30%
No. Replications	88	92	91	93
No. Studies	16	16	16	16

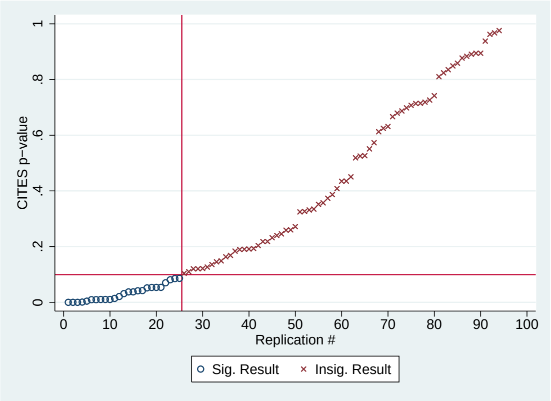
CITES

One possible concern is that the GATT/WTO regime truly does have a causal effect on a variety of dependent variables, perhaps in ways that we have failed to imagine. While we believe this is highly unlikely, our results obtain even when we use a more conservative replication approach. We also replicated all of the analysis conducted above, only we used the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) treaty instead of the GATT/WTO. CITES is a convention designed to safeguard certain species from over-exploitation. CITES went into force in 1975 and 179 countries are Parties to the convention.

The CITES treaty is very close to a "true placebo" test. It has virtually no theoretical link to any of the dependent variables analyzed. Its rules only govern a minuscule percentage of global trade and compliance with those rules is inconsistent at best. It is extremely unlikely that CITES membership has any causal effect on the dependent variables we analyze.

Table 3 replicates the results from the first table above. The false positive rates, 27%, are only slightly lower than those found above. In the reduced replications, the false positive rate was 35% and *rose* to 36% when we added country fixed effects. Time fixes and matching only lowered the

Figure 4: P-values for Effect of CITES on Irrelevant DV's



false positive rate to 27% and 22% respectively.

Table 3: False Positive Rates for Replications, CITES Variable

	Specification				
	Orig.	Reduced	Country FE	Splines/Country Trend	Matching
False Pos. Rate	27%	35%	36%	27%	22%
No. Replications	94	94	91	94	90
No. Studies	16	16	16	16	16

The same problem found above, where fixes remove some false positives while also creating new ones, is again present. Figure 4, Figure 5, and Figure 6 replicate the same series of figures that we presented in the GATT/WTO replications. Figure 4 shows the p-values from the original replications, using the CITES variable. Figure 5 and Figure 6 retain the same ordering of studies from Figure 4 and show the new p-values. Country fixed effects make the CITES variable insignificant in 4 of the original replications, yet make the CITES variable significant in 12 replications where it was insignificant before. Matching fares slightly better, removing 13 false positives, but creating 9 new ones.

Combinations of fixes again fail to lower the false positive rate, as shown in Table 4, which repeats the same series of specifications as in Table 2. The false positive rate is lowest when using country specific linear trends in isolation, but is still too high (24%). Adding country and/or year fixed effects again raises the false positive rates back above 29%, even reaching 37% in the replications with year fixed effects.

Figure 5: P-values for Effect of WTO on Irrelevant DV's, Fixed Effects

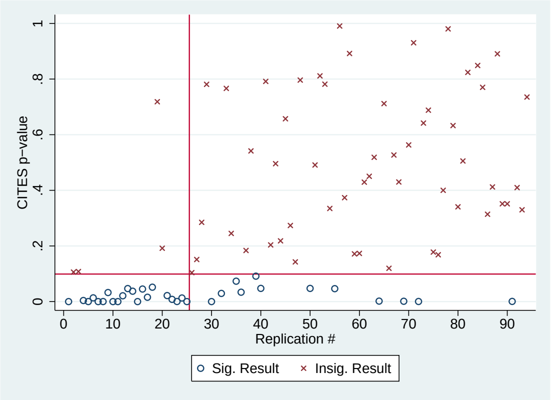


Figure 6: P-values for Effect of WTO on Irrelevent DV's, Matching

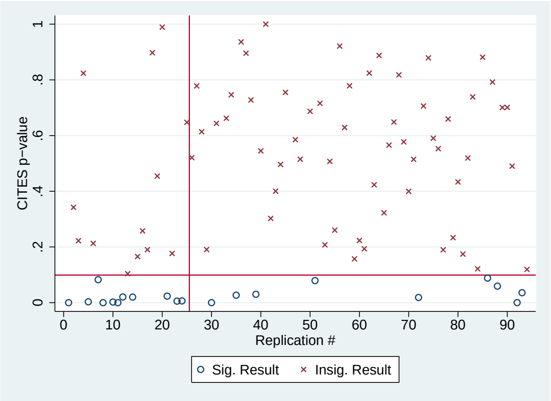


Table 4: False Positive Rates for Replications with Multiple "Fixes," CITES Variable

	Specification			
	Cty. Trends	Cty. Trends + Cty. FE	Year FE	Cty. and Year FE
False Pos. Rate	24%	35%	37%	29%
No. Replications	88	92	91	93
No. Studies	16	16	16	16

Simulations and a Generic Data Generating Process

The preceding section established that false positives are likely a problem. This section generates intuition using a controlled environment where the true data-generating process (DGP) is known. We first describe a general DGP that is theoretically grounded in our understanding of treaties and compliance. This general DGP accommodates several possible types of unobservables. We describe each type mathematically and motivate them with real-world arguments.

We then describe a simpler DGP and conduct Monte Carlo simulations to demonstrate two key points. First, the false positives problem that we observed in the replication exercise was not an artifact of the studies we chose or the way we replicated the authors' results. The DGP *explicitly sets the effect of a treaty on compliance to be zero,* so any significant results we recover from the simulations are *by definition* false positives. Even in situations where the DGP is carefully controlled, commonly used approaches are prone to generate false positives.

Second, the simulations demonstrate how, as we saw in the replication exercise, using a fix for one problem can exacerbate others. When researchers choose their empirical strategy to account for one type of unobservable, they can often make things worse if other types of unobservables are present. We describe a "law of second best solutions." In economics, this term refers to situations where fixing one, but not all, market imperfections, can decrease aggregate welfare. A similar phenomenon occurs here. If the empirical model can't account for *all* types of unobservables, then fixing some but not all aspects of the problem may make the results more susceptible to false positives.

Data-Generating Process with Types of Unobservables

As in the previous sections, let X_{it} be a vector of observable characteristics of country i in year t which potentially affect both the decision to ratify a treaty and its decision to comply. Let r_{it} be an indicator variable that equals 1 if country i ratified the treaty in year t and zero otherwise. The "1" denotes an indicator function, where the variable takes on a value of 1 if the condition in parenthesis is met. We call Equation 2 the ratification equation and Equation 3 the compliance equation.

$$r_{it} = 1(X_{it}B + u_{it}^r > 0) \tag{2}$$

$$c_{it} = X_{it}\beta + \alpha r_{it} + u_{it}^c \tag{3}$$

Unobservables could be like the following composite disturbances for the ratification and compliance equations, where disturbances are broken down into different "types." For each component, we use the superscripts r and c to indicate whether the observable enters the ratification or compliance equation.

Unobs. in ratification equation

$$\begin{aligned} u_{it}^r &= \mu_i^r + \delta_t^r + \gamma_i^r t + e_{it}^r \\ \mu_i^r &\sim N(m^r, \sigma_{r1}^2) \\ \delta_t^r &\sim N(d^r, \sigma_{r2}^2) \\ \gamma_i^r &\sim N(g^r, \sigma_{r3}^2) \\ e_{it}^r &\sim N(e^r, \sigma_{r4}^2) \end{aligned}$$

Unobs. in compliance equation

$$\begin{aligned} u_{it}^c &= \mu_i^c + \delta_t^c + \gamma_i^c t + e_{it}^c \\ \mu_i^c &\sim N(m^c, \sigma_{c1}^2) \\ \delta_t^c &\sim N(d^c, \sigma_{c2}^2) \\ \gamma_{it}^c &\sim N(g^c, \sigma_{c3}^2) \\ e_{it}^c &\sim N(e^c, \sigma_{c4}^2) \end{aligned}$$

Bias in estimates of α arise from the correlation between each type of unobservable across the ratification and compliance equations. We characterize the correlations between each type of unobservable in the ratification and compliance equations as follows:

$$\begin{aligned} cov(\mu^r, \mu^c) &= \rho_1 \\ cov(\delta^r, \delta^c) &= \rho_2 \\ cov(\gamma^r, \gamma^c) &= \rho_3 \\ cov(e^r, e^c) &= \rho_4 \end{aligned}$$

In these composite disturbances, there are three distinct types of unobservables. μ_i represents a country-specific unobservable. In many contexts, we would expect this type of unobservable. Consider the difficulty in assessing whether membership in the GATT/WTO causes countries to trade more. There are many country-specific factors that affect whether/when a country joins the GATT/WTO and the amount they trade. For example, larger, more globalized and more prominent countries were among the GATT founding members. And it is entirely plausible that these countries also tend to trade more. If left unaccounted for, these factors bias us in favor of finding that GATT/WTO membership increases trade, even if it truly has no effect. Some of these factors might be easy to observe and account for. If country size is the confounding factor, then researchers could measure and control for a country's GDP in some way. Level of globalization or global prominence might be harder to observe.

δ_t represents a year-specific component. This component describes factors which vary over time, affecting ratification and compliance. To continue the GATT/WTO and trade example from above, there are many candidates. Shipping costs decreased over time which could encourage countries to join the GATT/WTO and also to trade more. Consumers may, increasingly over time, love a variety of international goods coming from many different suppliers which could influence GATT/WTO membership and trade. Again, the presence of these types of year-specific unobservables or global trends bias estimates of the effects of the GATT/WTO on trade upwards. Shipping costs may be easy to observe and control for, while consumer tastes may not.

γ_{it} represents a country-specific time trend. Countries may be on different trajectories with respect to ratification and compliance. For example, new (and new new) trade theories suggest that firms or countries can benefit from economies of scale of production, which might increase their market shares or drive out competitors. It is plausible that early ratifiers of the GATT/WTO were also the types of countries who could benefit from economies of scale, which would make the trend in their amount of trade more steeply sloped over time. These types of factors may be particularly difficult to observe and measure, since they may be based on features of the world further back in time and since they might rely on relative values of certain variables.

More complex types of unobservables are certainly possible. The DGP above has linear country-specific trends. There could be higher-order trending. Country specific unobservables could be common to a region or area, etc. Our point is not that we have exhausted the features of the real world's DGP, but rather that the problem of unobservables is multifaceted. There are many theoretically plausible types of unobservables which make estimating the effect of a treaty on compliance difficult.

Simpler Data-Generating Process and the Law of Second Best

Our two main results from above, (1) that many fixes do not fix the problem of false positives and (2) fixes can help or hurt, obtain even with simulations from a simpler, known DGP.

The simpler DGP that we use consists of the following system of equations:

$$\begin{aligned} & r_{it} = 1(x_{it} + u^r_{it} > 0) \\ & c_{it} = x_{it} + \alpha r_{it} + u^c_{it}, \\ & \text{or} \\ & c_{it} = 1(x_{it} + \alpha r_{it} + u^c_{it} > 0) \end{aligned} \quad ,$$

where $x_{it} \sim N(0, \sigma^2)$, $\alpha = 0$, and u^r_{it} and u^c_{it} are composite random disturbances. Note that the DGP generates a continuous and binary compliance variable, which makes it more flexible than the equations described in preceding sections.

The two simplifications for this DGP are as follows. First, we include only one covariate, x_{it}, which affects both membership and compliance. Second, we limit the "types" of selection on unobservables that are present. Since we only need two sources of correlation across disturbances to demonstrate the basic problem, we generate our disturbances as:

$$\begin{aligned} u^r_{it} &= \sqrt{.5}\mu^r_i + \sqrt{.5}\delta^r_t \\ u^c_{it} &= \sqrt{.5}\mu^c_i + \sqrt{.5}\delta^c_t \end{aligned}$$

Each disturbance has two components, a unit and period-specific effect. These are jointly normally distributed as:

$$\begin{bmatrix} \mu^r_i \\ \mu^c_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_\mu \\ \rho_\mu & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} \delta^r_t \\ \delta^c_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_\delta \\ \rho_\delta & 1 \end{bmatrix} \right)$$

It follows that the composite disturbances are also jointly normally distributed

$$\begin{bmatrix} u_{it}^r \\ u_{it}^c \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

and that the covariance/correlation can be decomposed as

$$\rho = .5\rho_\mu + .5\rho_\delta$$

where ρ_μ represents between-unit contribution to the overall covariance and ρ_δ represents the within-unit contribution to the overall covariance.

For our simulations, we set the number of units or countries to be $N = 100$ and the number of years to be $T = 30$. These values are similar to those observed in the replication exercise above. We set the variance of our observable covariate equal to one ($\sigma^2 = 1$), which implies that x_{it} accounts for half of the variance in our continuous compliance and latent compliance outcomes.

We consider results from four cases of replications. The cases differ from one another in two ways. First, moving from Case 1 to Case 4, we gradually increase the overall covariance between the ratification disturbance term and the compliance disturbance term from $\rho = .25$ (Case 1) to $\rho = .75$ (Case 4). In other words, the overall problem of selection on unobservables gradually gets worse.

The cases also differ in the type of correlation across disturbances. In our first two cases, all of the covariance between ratification and compliance disturbances is attributable to within-unit variance caused by our period effects. In our third and fourth cases, this covariance is attributable to both within- and between-unit variance in the unobservables. In other words, the first two cases involve only one type of selection on unobservables, and the second two cases involve two sources.

For our continuous compliance experiments, we evaluated the performance of three approaches: OLS without any fixed effects ("do nothing"), unit fixed-effects, and matching estimators, as in Table 5. We use panel-corrected standard errors with our OLS and fixed-effects estimators. For our binary compliance experiments, we evaluated the basic logit, conditional logit, and matching estimators, as in Table 6. We used time-period clustered and panel-bootstrapped standard errors for our logit and conditional logit estimators respectively. The matching estimator is the same as in the replications. Evaluations are based on 1,000 trials.

We expect two trends in the results. First, the false-positive performance of the "do-nothing" estimators should deteriorate across our cases as we move from low to high covariance between

Table 5: MCs for Continuous DVs ($N = 100$, $T = 30$, $\sigma = 1$, 1000 trials)

	$\rho_\mu = 0$	$\rho_\mu = 0$	$\rho_\mu = .5$	$\rho_\mu = .75$
	$\rho_\delta = .5$	$\rho_\delta = .75$	$\rho_\delta = .5$	$\rho_\delta = .75$
	$\rho = .25$	$\rho = .375$	$\rho = .5$	$\rho = .75$
OLS				
Mean($\hat{\alpha}$)	.393	.593	.808	1.219
S.d.($\hat{\alpha}$)	.171	.174	.16	.148
Mean(s.e.($\hat{\alpha}$))	.143	.134	.141	.130
Overconfidence	1.196	1.299	1.135	1.138
False Positive Rate	75.9%	98.9%	99.9%	100%
Fixed Effects				
Mean($\hat{\alpha}$)	.533	.799	.532	.803
S.d.($\hat{\alpha}$)	.192	.187	.195	.189
Mean(s.e.($\hat{\alpha}$))	.183	.164	.183	.163
Overconfidence	1.049	1.140	1.066	1.160
False Positive Rate	84.1%	99.8%	83.7%	99.9%
Matching				
Mean($\hat{\alpha}$)	.443	.659	.899	1.338
S.d.($\hat{\alpha}$)	.213	.209	.195	.165
Mean(s.e.($\hat{\alpha}$))	.111	.106	.101	.086
Overconfidence	1.919	1.972	1.931	1.919
False Positive Rate	85.4%	98%	99.9%	100%

the ratification and compliance disturbances. Second, the relative performance of our fixed-effects estimators should improve in our high covariance cases where some of the overall covariance is attributable to unit effects, but deteriorate when this is not the case.

In the case of the unit fixed effects estimator, this arises because of simultaneity bias. Ratification is endogenous if it covaries with the disturbance in the compliance equation. Fixed-effects estimators potentially reduce this covariance, but they also reduce the variance in the ratification decisions that is leveraged to estimate their causal effects on compliance. The bias in the estimated treatment effect depends on both of these. The simultaneity bias increases with the strength of the covariance between ratification decisions and the unobservable determinants of compliance, but it decreases as the variance in ratification decisions increases. The first-best solution is to eliminate *all* of the spurious sources of covariance between ratification and compliance. If this can be done, the causal effect is identified. However, if only some of these sources can be eliminated, the estimator's performance can be worse than doing nothing. In fact, the second-best solution may be to do nothing. Plumper and Troeger (2013) make a similar point, finding that unit-fixed effects strategies may be worse than pooled strategies in the presence of unobserved trending. Clarke (2005) and Clarke (2009) also find similar results when describing the effect of control variables.

Table 6: MCs for Binary DVs ($N = 100$, $T = 30$, $\sigma = 1$, 1000 trials)

	$\rho_\mu = 0$	$\rho_\mu = 0$	$\rho_\mu = .5$	$\rho_\mu = .75$
	$\rho_\delta = .5$	$\rho_\delta = .75$	$\rho_\delta = .5$	$\rho_\delta = .75$
	$\rho = .25$	$\rho = .375$	$\rho = .5$	$\rho = .75$
Logit				
Mean($\hat{\alpha}$)	.662	1.034	1.47	2.569
$\Delta\,\Pr(c_{it} = 1 \vert R_{it} = 1)$	.160	.238	.313	.429
S.d.($\hat{\alpha}$)	.292	.296	.287	.275
Mean(s.e.($\hat{\alpha}$))	.251	.242	.251	.241
Overconfidence	1.163	1.223	1.143	1.141
False Positive Rate	71.2%	97.3%	99.9%	100%
Conditional Logit				
Mean($\hat{\alpha}$)	1.594	2.83	1.579	2.772
$\Delta\,\Pr(c_{it} = 1 \vert R_{it} = 1)$	.331	.444	.329	.441
S.d.($\hat{\alpha}$)	.565	.577	.568	.549
Mean(s.e.($\hat{\alpha}$))	.158	.209	.144	.165
Overconfidence	3.576	2.761	3.944	3.327
False Positive Rate	98.5%	100%	98.7%	100%
Matching				
Mean($\hat{\alpha}$)	.124	.194	.281	.487
S.d.($\hat{\alpha}$)	.069	.072	.075	.074
Mean(s.e.($\hat{\alpha}$))	.048	.048	.046	.041
Overconfidence	1.438	1.5	1.630	1.805
False Positive Rate	63%	90.2%	99.5%	100%

The rows marked $\Delta\,\Pr(c_{it} = 1 | r_{it} = 1)$ denote the substantive effects of ratification in terms of first differences. They show the change in the probability that compliance = 1 for ratifiers compared to non-ratifiers.

Including an additional control variable could increase or decrease bias in the resulting estimates of interest.[12]

We find both of these patterns in the results. In both Table 5 (continuous compliance) and Table 6 (binary compliance) the performance of the do-nothing OLS and matching estimators gets progressively worse as we move from Case 1 to Case 4. With respect to fixed effects, as expected, we see that its performance is worse than OLS when none of ρ is attributable to between-unit covariance ($\rho_\mu = 0$) and better when half of ρ is attributable to between-unit covariance ($\rho_\mu = .5$). We are not interested in identifying the exact threshold at which fixed-effects begins to outperform OLS. This is highly dependent on the nature of the DGP. The basic point, however, is generalizable: if the fixed-effects strategy does little to address the covariance between the unobserved disturbances that determine both ratification and compliance, but does reduce significantly the variance in the unobserved disturbance that determines ratification, it will make the simultaneity bias worse.

[12] For more general discussions of a similar phenomenon, see Pearle (2000) or Spirtes, Glymour and Scheines (1993).

The high false positive rates of the matching estimator further support the argument made above that, even when the researcher can achieve balance on observables, this does not insulate against false positives resulting from imbalance on unobservables. In the Monte Carlo simulations we do very well in achieving balance on observables. Yet, we still have high false positive rates. This further confirms that our results in the replications sections above are not artifacts of failure to achieve balance on observables or failure to use a particular matching algorithm.

Sensitivity Tests

The types of unobservables which affect ratification and outcomes like compliance in observational data are likely to be complex and multifaceted. All applied empirical work makes assumptions about these unobservables and produces estimates which are influenced by those assumptions. When those assumptions do not match the "true" DGP, we risk producing biased estimates. This is particularly daunting since assumptions about unobservables are inherently untestable.

Fortunately, sensitivity tests condition inference even when the true nature of unobservables is unknown. These tests generate leverage by asking: "how severe would the selection on unobservables problem have to be to make my estimated result a false positive?" The two approaches we describe here differ in how they benchmark "severity."[13]

The first approach, from Altonji, Elder and Taber (2005), gives answers of the form "selection on unobservables would need to be this severe, *relative to selection on observables*, to drive our estimated effect to zero." The italicized term highlights the benchmark of the test. The researcher benchmarks the result from this sensitivity test against what she thinks she knows about the *selection process*. We explain this approach in greater detail below, and to the best of our knowledge, this is one of its first political science applications.[14]

The second approach gives answers of the form "the effect of unobservables on compliance would need to be this severe, *relative to the effect of observables on compliance*, to drive our estimated effect to zero." This approach benchmarks the results from the sensitivity test against what the researcher thinks she knows about *compliance process*. Several related approaches to sensitivity fall into this category, which is often most closely associated with the work of (Imbens, 2003).

[13] Note, our goal is not an exhaustive characterization of all sensitivity tests. The two categories presented here are meant as a useful grouping of several prominent approaches. We do not cover approaches based on bounds, e.g. Manski (1990); Mebane and Poast (2013).

[14] The only other applications we are aware of are Besley and Reynal-Querol (2014), de Figueiredo and Edwards (2007), and Fair et al. (2013).

Methodological work on sensitivity testing is a vibrant field.[15] Our contribution in this section is for the applied researcher. By comparing and contrasting two real-world applications, we give practical advice on when and how to use each approach. We show how the power of each test is determined by the researcher's prior theoretical knowledge. If the researcher knows more about the selection process, the first approach is more powerful. If she knows more about the compliance process, the second approach is more powerful. To further facilitate applied research, the appendix gives step-by-step descriptions of the sensitivity analysis with accompanying Stata code. Despite methodological advances, sensitivity analysis is not commonly used in applied political science research. Our hope is that the advice and demonstrations here will help integrate more sensitivity analysis into applied situations in which selection on unobservables is potentially a problem.

Benchmark 1: Selection on Observables

The Altonji, Elder and Taber (2005) approach leverages the idea that, if unobservables have only a weak effect on ratification, then the researcher doesn't need to worry as much about bias resulting from selection on unobservables. If the effect is strong, then she does. To assess this, the test asks: how much stronger does selection on unobservables need to be, *relative to selection on observables*, in order to imply that there is no effect of ratification on compliance?

If, using this approach, the researcher finds that the strength of unobservables for explaining ratification has to be many times stronger than the effect of observables on ratification, then she can be confident in her estimated effects. If she finds that the strength of unobservables need only be a fraction of the strength of observables, she should be worried. The quantity of interest generated by this approach is a ratio: the ratio of strength of unobservables, relative to the strength of observables, which would drive the estimated effect of ratification to zero.

To calculate this ratio, we first need an expression for the bias in the estimated effect of ratification resulting from selection on unobservables.[16] This bias can be expressed as:

$$\text{plim } \hat{\alpha} = \alpha + \frac{\text{var}(r_{it})}{\text{var}(u^r_{it})} \left[E[u^c_{it}|r_{it}=1] - E[u^c_{it}|r_{it}=0] \right],$$

As before, r_{it} describes whether country i has ratified in or before year t. X is a matrix containing the observables. c_{it} describes whether country i complied in year t. u^r_{it} are the disturbances from a regression of ratification on the observables. $E[u^c_{it}|r_{it}=1] - E[u^c_{it}|r_{it}=0]$ describes the

[15] For two recent advances, see (Blackwell, 2014) on confounding functions and (Imai, Keele and Yamamoto, 2010; Imai et al., 2011) on mediation analysis.

[16] For a formal derivation, see appendix.

degree of selection on unobservables. It is the shift in the distribution of unobservables affecting compliance when comparing ratifiers and non-ratifiers. The term $\frac{var(r_{it})}{var(u^r_{it})}$ is necessary to adjust the bias expression by making treatment and the observables orthogonal. Under the null hypothesis of no ratification effect, i.e. $\alpha = 0$, this expression implies Equation 4

$$E[u^c_{it}|r_{it} = 1] - E[u^c_{it}|r_{it} = 0] = \hat{\alpha}\frac{\text{var}(u^r_{it})}{\text{var}(r_{it})} \tag{4}$$

The left hand side represents degree of selection on unobservables necessary to explain all of the estimated ratification effect. This quantity is not knowable. The innovation Altonji, Elder and Taber (2005) is to use "the degree of selection on observables as a guide to the degree of selection on unobservables (Altonji, Elder and Taber, 2005, p. 153)." Formally, we assume that the standardized selection on unobservables is the same as the standardized selection on observables.[17]

$$\frac{E[u^c_{it}|r_{it} = 1] - E[u^c_{it}|r_{it} = 0]}{\text{var}(u^c_{it})} = \frac{E[X'\gamma|r_{it} = 1] - E[X'\gamma|r_{it} = 0]}{\text{var}(X'\gamma)}. \tag{5}$$

γ is the vector of coefficients resulting from a regression of c on X where α is constrained to be zero. u^c_{it} comes from regressing c on X and the disturbances recovered from the ratification equation, $\hat{u}^r_{it}$. Substituting Equation 5 into Equation 4 gives us the desired ratio of selection on unobservables to observables, necessary to drive the effect of ratification to zero:

$$\frac{\hat{\alpha}\ \text{var}(u^r_{it})\text{var}(X'\gamma)}{\text{var}(r_{it})\text{var}(u^c_{it})(E[X'\gamma|r = 1] - E[X'\gamma|r = 0])} \tag{6}$$

If this ratio is high then the degree of selection on unobservables, relative to selection on observables, must be high in order to explain away the ratification effect. If the researcher believes that the observables are strong predictors of ratification, then she might believe that a high ratio is implausible, and therefore, that her result is robust. If the ratio is low, then selection on unobservables need only be weak to explain away the ratification effect.

In practice, calculating this ratio involves recovering quantities from simple regressions. First, the residuals u^r_{it} are recovered from regressing ratification on the observables. Second, regressing compliance on those residuals and the observables yields $\hat{\alpha}$. Third, estimating a constrained equa-

[17] Qualitatively, this is equivalent to assuming that, from a set of covariates which potentially affect ratification and compliance, we have chosen randomly. For a more formal description of this assumption, see Altonji, Elder and Taber (2002). To the extent that covariates are chosen to minimize omitted variable bias in the estimated effect of ratification, this condition will be conservative. However, if there are any covariates that are orthogonal to ratification, their inclusion in equation (5) will reduce the degree to which the Altonji et al. benchmark is conservative. Practitioners should choose the covariates for these sensitivity tests carefully.

tion where the effect of ratification is constrained to equal zero (e.g. regressing compliance on the observables, but not ratification) yields u^r_{it} and γ. The appendix describes each step in greater detail.

Benchmark 2: Explanatory Power of Observables

A second approach uses the explanatory power of observable covariates as a benchmark for quantifying the problem of unobservables. This approach is better known in political science, so our description is more brief.[18] This approach uses point-biserial correlation coefficients, which describe correlation when one variable is dichotomous (ratification) and the other is continuous (compliance).[19]

The point-biserial correlation coefficient between the binary ratification variable and the continuous unobservable determinant of compliance is

$$r = \frac{E[u^c_{it}|r_{it} = 1] - E[u^c_{it}|r_{it} = 0]}{[\text{var}(u^c_{it})]^{\frac{1}{2}}} \left[\frac{n_1 n_0}{n^2}\right]^{\frac{1}{2}},$$

where n is the total number of observations, n_1 is the number of observations under ratification, and n_0 is the number of observations not under ratification. We then calculate the proportion of the variance in outcomes that would be explained by the implied imbalance in unobservables across the ratification and non-ratification groups under the null hypothesis, $\alpha = 0$:

$$r^2 = \frac{\hat{\alpha}^2 [\text{var}(u^r_{it})]^2}{\text{var}(r_{it})\text{var}(u^c_{it})}.$$

This quantity can be compared to the explanatory power of observable covariates using their partial coefficients of determination. If the imbalance in unobservables under the null hypothesis would have substantially more explanatory power than powerful covariates, we conclude that such an imbalance is unlikely and that the estimated effect of ratification is robust. If the imbalance in unobservables under the null hypothesis would have relatively low explanatory power when compared with the observable covariates, we would conclude that the effect is sensitive to selection on unobservables.

[18] The full details are described in the appendix. For another description, see Clarke (2009) and Clarke (2005).

[19] Blackwell (2014) and Imai, Keele and Yamamoto (2010); Imai et al. (2011) also use R^2's or the relevant coefficient of determination.

Which Benchmark?

Our advice regarding the appropriate benchmark is simple. When the researcher knows a lot about the ratification process, it makes sense to benchmark the implied imbalance in unobservables against the imbalance in the relevant observables (the first approach). When the researcher does not know a lot about the ratification process or knows more about the compliance process it is better to benchmark against the explanatory power of covariates, their ability to explain compliance (the second approach).

To demonstrate this in practice, we use two replications from Gerring, Thacker and Moreno (2005) which were included in the replications above. Gerring et al. develop a theory of centripetalism in which they argue that institutions which centralize political authority and promote inclusion lead to good governance. They operationalize centripetal governance – their explanatory variable of interest – and test their theory using regression analysis. Among their results, Gerring et al. find that centripetalism is associated with higher government revenues from taxes and tariffs and lower illiteracy rates. We concentrate on these two outcomes.

In our replications, we found a positive relationship between GATT/WTO membership and government revenues and found a positive relationship between CITES membership and illiteracy rates. We – the researchers – suspect that both results are false positives. Fortunately, we can use our knowledge of ratification of the WTO to assess the first false positive. However, we do not know as much about ratification of the CITES treaty, so we must use our knowledge of explanations for literacy to assess the second false positive.

Government Revenue False Positive

The government revenue replications suggest that WTO membership increases government revenue as a share of GDP by 3.69%, *ceteris paribus*. The estimated coefficient is statistically significant (t-statistic of 6.96). The result is robust to including country fixed effects and the matching approach.

This positive relationship is almost certainly spurious. The WTO explicitly limits tariff barriers. The government revenue data include tariffs as a source of revenue. If there is any effect of WTO membership on government revenue, it should be negative since membership requires countries to lower their tariffs, but does not require other changes to their tax policy.

To assess the likelihood of a false positive relationship, we use the first benchmark, because we feel more confident in our knowledge of the ratification process for the WTO. Recent work from Davis and Wilf (2011) argues that political and economic variables affect which countries join the GATT/WTO. For political explanations, they argue that the GATT/WTO is a "like-minded club" which admits new members who share important characteristics with existing members. They find that a country's level of democracy, measured by Polity scores, robustly predicts its time to GATT/WTO application. For economic explanations, they argue that large-economy countries benefit from making commitments that prevent them from using tariffs to extract rents, and are therefore more likely to apply for GATT/WTO membership. They find that both GDP and GDP per capita explain time to application.

Fortunately, several variables in the Gerring et al study measure similar quantities to those which Davis and Wilf identify as important determinants of ratification. From the Gerring et al study, we select Centripetalism, Democracy stock, GDP per capita, and Population from the set of covariates for our sensitivity analysis. We also include Oil production since, as Davis and Wilf note, oil is not governed by the trade regime, which may discourage membership among oil exporters. We use these variables to benchmark the degree of selection on observables.

We first calculate the implied imbalance in unobservables under the null hypothesis that WTO membership has no effect on government revenue, which is 0.08.[20] We use ˆ to denote estimates recovered from particular regressions.

$$\hat{\alpha}\frac{\widehat{\text{var}}(\hat{u}^r_{it})}{\widehat{\text{var}}(r_{it})\widehat{\text{var}}(\hat{u}^c_{it})} = 3.69\frac{.11}{(.17)(30.49)} = .08.$$

Using the five covariates which theoretically affect WTO membership (centripetalism, democracy stock, GDP per capita, population and oil production), the standardized imbalance in observables is .18,

$$\left[\hat{E}[X'\hat{\gamma}|r_{it}=1] - \hat{E}[X'\hat{\gamma}|r_{it}=0]\right]\Big/\hat{\text{var}}(X'\hat{\gamma}) = [3.85 - 1.16]/14.73 = .18.$$

Under the null hypothesis of no ratification effect, selection on unobservables would only have to be $\frac{0.08}{0.18} = 0.44$ as strong as selection on the relevant observables. This seems plausible. The five variables we identified have a theoretical relationship with WTO membership, but they are unlikely to explain *all* of WTO membership. It is very possible that one of more unobservables are approximately half as strong at explaining WTO membership as the observables we used here. This suggests that the relationship between WTO membership and government revenue is likely a

[20] This is the expression from Equation 4, only standardized by multiplying it by $\frac{1}{\text{var}(u^c_{it})}$, to better present the ratio as a fraction.

false positive. The approach helps us ground our skepticism about this result because of our ability to say something theoretical about the likely existence unobservables affecting ratification, relative to observables.

To see how this approach breaks down when the researcher lacks this theoretical knowledge, consider what happens if we treat *all* of the covariates in the Gerring et. al. study as possible observables which affect ratification, even those which lack a theoretical link to WTO ratification. When we include these variables, the linear prediction, $X'\gamma$, now contains more noise that is orthogonal to selection. This increases $var(X'\gamma)$, but has a minimal effect on $E[X'\gamma|r_{it} = 1] - E[X'\gamma|r_{it} = 0]$.

This is indeed what happens. The standardized difference in observables across ratifiers and non-ratifiers when we include all the covariates is .05,

$$\left[\hat{E}[X'\hat{\gamma}|r_{it} = 1] - \hat{E}[X'\hat{\gamma}|r_{it} = 0]\right] \Big/ \widehat{\text{var}}(X'\hat{\gamma}) = [23.00 - 20.67]/42.38 = .05.$$

Thus, under the null hypothesis of no treatment effect, selection on unobservables would have to be 1.6 times stronger that selection on observables. This is much less likely than our previous finding of .44. As a result, the sensitivity test appears to have gotten more strict and we're more likely to think that the estimated effect of ratification is a true positive. Careful theoretical attention must be paid to choose which observables to use. This approach has the most power when the researcher is most theoretically confident in the observables chosen, and its power weakens as confidence in the theoretical relationship decreases.

Literacy False Positive

Next, we turn to the Gerring et al. illiteracy regression. When we include our CITES variable in their literacy regression, we find that ratifying CITES causes the illiteracy rate among adults to rise by approximately 16%, *ceteris paribus*. The estimate is statistically significant (t-statistic of 2.57). This result is robust to including country fixed effects and using the matching approach. This is almost certainly a false positive result. We are unaware of any theoretical relationship between a minor environmental treaty and literacy rates.

The Altonji et. al. approach is not useful here. We do not have other work or prior knowledge to allow us leverage over the causes of CITES ratification. As a result, even if we calculated the ratio of selection on unobservables relative to selection on observables necessary to drive the effect of ratification on literacy to zero, we would have no meaningful way to assess this ratio.

Fortunately, we can use the second benchmark, leveraging what we know about explanations for literacy– namely, the theoretical and empirical relationships between Gerring et. al.'s covariates and literacy. The partial coefficients of determination, describing the effect of Gerring. et. al.'s covariates on literacy, range from 0 to 0.36 with a median of .018. Under the null hypothesis of no treatment effect, the implied difference in unobservables across the CITES and non-CITES countries would have a partial coefficient of determination very close to zero,

$$r^2 = \frac{\hat{\alpha}^2[\widehat{\text{var}}(\hat{u}^r_{it})]^2}{\widehat{\text{var}}(r_{it})\widehat{\text{var}}(\hat{u}^c_{it})} = \frac{.16^2 \times .11^2}{.25 \times 2.49} = 4.98 \times 10^{-4}.$$

This is lower than all but 3 of the partial coefficients of determination for the variables in Gerring et al.'s model. A very small difference in unobservables across the CITES and non-CITES groups could explain the entire estimated treatment effect. Since it is very plausible that such a difference exists, we can confidently say that this is a false positive.

Conclusions

This paper has covered a lot of ground. We conclude with the following remarks.

First, recognizing the problem is inherently important. There are strong theoretical reasons to expect that unobservables affect ratification and compliance. This generates false positives, where we mistakenly conclude that certain institutions cause compliance. As shown with a replication exercise using existing work and with Monte Carlo simulations, this problem is potentially severe and multifaceted. We found false positive rates generally around 34%, which is much higher than would be tolerated by conventional assessments of statistical significance.

Second, there is no universal "fix." Neither matching nor fixed effects nor combinations of various approaches are likely to resolve this problem without strong prior theoretical knowledge about the underlying data generating process. This problem is exacerbated by "the law of second best" which describes how addressing one aspect of the selection on unobservables problem, without addressing all aspects, can make the problem worse. Under different conditions, fixes can raise or lower false positive rates, and these conditions are not generally things for which the researcher has strong prior theoretical knowledge. We demonstrated the law of second best, and confirmed our findings from the replication experiment, using carefully controlled Monte Carlo simulations.

Third, theoretically informed sensitivity analysis is a powerful tool for assessing whether a particular result is a false positive. All existing approaches and fixes rely on untestable assumptions.

Sensitivity analysis allows the researcher to assess how sensitive her estimates are to alternative assumptions about the severity of the selection on unobservables problem. When she has strong theoretical knowledge about ratification, she can benchmark her assessment of unobservables relative to the effect of observables on ratification. When she instead has strong theoretical knowledge about compliance, she can benchmark her assessment relative to the explanatory power of observables on compliance. Ultimately, the persuasiveness of these approaches is founded on the researcher's theoretical knowledge against which she will benchmark her results.

Finally, our strongest emphasis is on the relationship between theoretical knowledge and empirical models. Each and every facet of the problem of false positives, its existence, severity, solution, and assessment, requires the researcher to think carefully about the underlying data generating process and what she theoretically believes about it. These beliefs hopefully are persuasive, based on logically consistent models of behavior, supported by ancillary data or experience, or commonly agreed upon. Because at each and every step, they are called upon. The search for one "fix" to the selection on unobservables problem or a fool-proof sensitivity test that does not require the researcher to carefully draw on her theoretical knowledge is quixotic. We hope that we have provided applied researchers with tools to leverage their theoretical knowledge in the face of selection on unobservables.

References

Altonji, Joseph G., Todd E. Elder and Christopher R. Taber. 2002. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." Manuscript. Dept. Econ., Northwestern Univ.

Altonji, Joseph G., Todd E. Elder and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1):151–184.

Bertrand, M., E. Duflo and S. Mullainathan. 2004. "How Much Should We Trust Difference-In-Differences Estimates?" *Quarterly Journal of Economics* 119(1):249–275.

Besley, Timothy and Marta Reynal-Querol. 2014. "The Legacy of Historical Conflict: Evidence from Africa." *American Political Science Review* 108:319–336.

Blackwell, Matthew. 2014. "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." *Political Analysis* 22(2):169–182.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341–352.

Clarke, Kevin A. 2009. "Return of the Phantom Menace: Omitted Variable Bias in Political Research." *Conflict Management and Peace Science* 26(1):46–66.

Davis, Christina and Meredith Wilf. 2011. "Joining the Club: Accession to the GATT/WTO." Working Paper, Princeton University.

de Figueiredo, Rui J. P., Jr. and Geoff Edwards. 2007. "Does Private Money Buy Public Policy? Campaign Contributions and Regulatory Outcomes in Telecommunications." *Journal of Economics and Management Strategy* 16(3):547 – 576.

Downs, George W., David M. Rocke and Peter N. Barsoom. 1996. "Is the Good News about Compliance Good News about Cooperation?" *International Organization* 50(3):379–406.

Fair, C. Christine, Patrick M. Kuhn, Neil Malholtra and Jacob N. Shapiro. 2013. "How Natural Disasters Affect Political Attitudes and Behavior: Evidence from the 2010-11 Pakistani Floods." Manuscript.

Gerring, John, Strom C. Thacker and Carola Moreno. 2005. "Centripetal Democratic Governance: A Theory and Global Inquiry." *The American Political Science Review* 99(4):pp. 567–581.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105:765–789.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):pp. 51–71.

Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *The American Economic Review* 93(2):pp. 126–132.

Leuven, E. and B. Siansei. 2003. "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.". Version 4.0.11.
URL: *http://ideas.repec.org/c/boc/bocode/s432001.html*

Lupu, Yonatan. 2013. "The Informative Power of Treaty Commitment: Using the Spatial Model to Address Selection Effects." *American Journal of Political Science* .

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review* 80(2):pp. 319–323.

Mebane, W.R. and Paul Poast. 2013. "Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data." *Political Analysis* 21(2):233–251. cited By (since 1996)2.

Pearle, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Plumper, Thomas and Vera E. Troeger. 2013. "Not So Harmless After All: Fixed Effects as Identification Strategy." EPSA Conference Paper.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1):41–55.

Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *The American Political Science Review* 94(4):819–835.

Simmons, Beth A. and Daniel J. Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *American Political Science Review* 99(04):623–631.

Spirtes, P., C. Glymour and R. Scheines. 1993. *Causation, Prediction, and Search*. Springer-Verlag, New York.

Von Stein, Jana. 2005. "Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance." *The American Political Science Review* 99(4):611–622.