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

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

September 18, 2014

*We appreciate helpful advice from Marc Busch, William Clark, Kristian Skrede Gleditsch, In Song Kim, Moritz Marbach, and Michael Miller. 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 international institutions on member state behavior is a difficult task since unobservable factors affect institutional membership and state behavior, which biases estimates in favor of finding a positive effect of institutions. We use a replication experiment of 94 specifications from 16 different studies to show the severity of this "false positives" 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, one of which has yet to be applied to political science and give guidance for the conditions under which each sensitivity approach can guard against the problem of false positives.

One of the most important questions in the study of international relations (IR) is: how are a country's policies different in a world in which they have joined an organization, ratified a treaty, or agreed to the rules of a particular institution, compared to a world in which they have not? 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 have been interested in whether member states change their policies to be in line with an institution's rules or prescriptions, i.e. compliance. Initially, the debate over this question was stark, with some scholars arguing that institutions rarely affect member state behavior and others arguing that they have significant effects. More recently, researchers have posited specific theoretical ways in which institutions might induce compliance and have empirically evaluated particular predictions using observational data.

At least since Downs, Rocke and Barsoom (1996) asked if the good news about compliance constituted good news about cooperation, we have known that it is difficult to use observational data to assess the effect of institutions on compliance. 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," is well understood in international relations applications and beyond. Its consequence is also well understood. Selection on unobservables most likely biases empirical findings regarding the effects of institutions in a positive direction. Even if the institution has no causal effect on compliance, selection on unobservables can result in "false positives," where estimates incorrectly suggest a positive effect of membership on compliance. When we observe a positive empirical relationship between institutional membership and compliance, we are left wondering whether this finding is the result of a causal relationship between membership and compliance, or whether the finding is really just an artifact of selection on unobservables.

Extant IR research uses a veritable smorgasbord of empirical models designed to address this problem, with unit fixed effects and matching estimators being the most popular. 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 to international political economy, e.g. a country's torture rate or whether it has a legislature. Using

identical models to the authors' original specifications, we assess whether a country's membership in the World Trade Organization (WTO) 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 false positives. The WTO has a statistically significant relationship with these theoretically-irrelevant dependent variables approximately 34% of the time. We also show how the most commonly used estimation approaches do not reduce these false positive rates, and in some instances, these fixes make the problem worse by creating new false positives where there were none before. These same results obtain even when using membership in the Convention on Trade in Endanged Species (CITES), which is an even more theoretically distant treaty than the WTO. This should give researchers significant pause regarding the degree to which extant approaches address the problem of selection on unobservables.

While the WTO and CITES replication exercise diagnoses the severity of the problem of false positives, the second section explains the problem in a controlled environment. We present a generic data generating process that highlights the key features of the selection on unobservables problem for IR researchers. Specifically, we show how 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 of unobservable is theoretically plausible and supported by arguments in existing literature. Yet each also implies different things 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. The simulations first confirm our results from the replications. Even when the DGP 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, which has not previously been used in political science, 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 entirety of the relationship found?" The second approach, from Imbens (2003), leverages the power of observables to explain outcomes to benchmark the likelihood of a false positive. Each approach has its strengths and weaknesses. Our goal in this section is to highlight their similarities and differences, and most importantly, to give guidance on when researchers should apply each approach. When the researcher has stronger theoretical, prior knowledge about the selection process, the first approach is more useful. When the researcher instead knows more about the outcome process, the second approach is more useful.

The key advice uniting all these sections is that researchers should recognize that all approaches require careful attention to their assumptions and how those assumptions compare to what the researcher thinks she knows about the particular situation. All estimation approaches and sensitivity tests require this, yet too often, applied research and methodological innovations treat a particular approach as a panacea for the problem of selection on unobservables. The search for one approach (estimator, sensitivity test, etc.) which can be applied to all situations is unrealistic. Only with careful attention paid to the relationship between assumptions and theoretical knowledge can researchers avoid the problem of false positives. Finally, it is worth noting that we have packaged these arguments in terms of IR research regarding the effects of institutions on behavior. Yet none of our arguments are idiosyncratic to that context. Selection on unobservables, the difficulty of assessing whether existing fixes work, and the futility of a search for a universal solution to this problem are characteristics of virtually all applied political science work.

The Problem of False Positives

A large body of research in international relations 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 is binary, and equals 1 if country i has joined that institution in or before year t, and zero otherwise. For the dependent variable, we are often interested in whether a country has adopted policies that are consistent with that institution's rules, i.e. 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, a large body of work 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, a large body of work asks 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]

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

Researchers generally are interested in estimating α, the effect of ratification on compliance. In estimating α, IR researchers face a familiar problem: the unobservables that affect ratification are positively correlated with the unobservables that affect compliance, which biases estimates of α 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]

Consider one possible unobservable: a country's *ex ante* costs to compliance. Moving from

[1] Compliance need not be binary. We describe it as binary here for simplicity. 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. We return to a more extensive discussion of these issues later.

[3] This problem has been described in great detail elsewhere, so we will eschew a lengthy review here. For some of the most well-known examples, see Simmons (2000); Simmons and Hopkins (2005); Von Stein (2005). For a more recent treatment, see Lupu (2013).

noncompliance to compliance potentially entails economic or political costs, which can be difficult to measure. Countries for whom compliance is easy also are the most likely to ratify. After all, if the point of the treaty is to increase the costs of noncompliance, then high-cost countries should be less likely to ratify and vice versa. If we observe some countries ratifying a treaty and complying with its obligations, and some countries neither ratifying nor making adjustments to their policy, we are tempted to conclude that the treaty caused compliance. But because of the possibility of selection on unobservables, we are left wondering whether compliance decisions are explained by the treaty or by the countries' *ex ante* costs of making certain policy adjustments.

To deal with this problem, extant work uses a large variety of "fixes." A broad class of approaches treats selection on unobservables as a type of omitted variable bias. If these unobservables are a type of unobserved across-unit heterogeneity, then standard panel data techniques, like unit fixed effects, can recover unbiased estimates of α, even in the presence of this type of unobservable. If unobservables arise due to trends over time that are common to all units or unit specific, then other standard techniques like time trends or splines may suffice, etc. Relatively more recently, many researchers have advocated matching algorithms as a solution to this problem. This motivation for matching in these contexts is largely 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.

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. The overall approach that we use to support this claim is to 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 have little to do with 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.

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 possible replications.

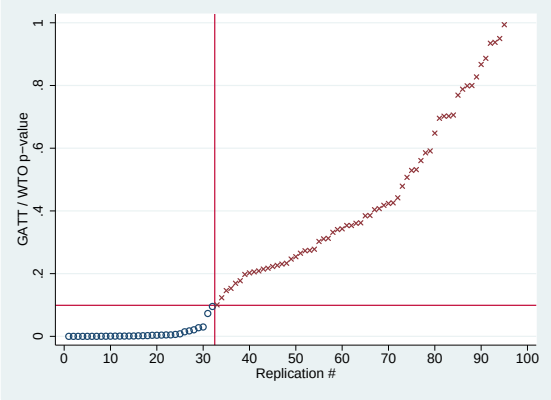
Baseline Replications

For the baseline set of replications, we used authors' exact original specifications. The only change we made was to simply add the WTO_{it} variable as an additional explanator. If the authors estimated a model expressing y_{it} as a function of observables X_{it}, then we estimated a model like

[4] We had to limit ourselves to studies where the authors provided replication materials online or upon request. We supplemented this set of studies 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. If anything, including these additional studies should make our test for the presence of false positives more conservative, because presumably, these studies are less susceptible to the problem of false positives. A full list of the studies and further details is available in the appendix.

[5] Practically speaking, we excluded all trade-related dependent variables, e.g. trade, tariffs, etc. While this is a subjective exercise, we tried to be conservative. We excluded any dependent variable for which we were aware of existing theoretical work linking WTO membership and that dependent variable. We also excluded categorical and ordered dependent variables. Since our goal will soon be to compare different strategies, such as fixed effects or matching, we excluded these dependent variables since it's very difficult to estimate comparable specifications. We have little reason to suspect that this exclusion biases our findings in any way.

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



the one expressed in Equation 2.

$$y_{it} = f(X_{it}\beta + \alpha WTO_{it} + e_{it}) \tag{2}$$

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 thus 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 did not. Blue O's in the bottom left correspond to "false positives." These are studies where the WTO appears to have 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 that are less than 0.10, meaning the false positive rate is approximately 34%. The false positive results are also far from "barely significant results." 30 of the replications have p-values that are less than 0.05. 25 of the replications have p-values that are less than 0.01.

The false positives are also not concentrated in just a few studies or just a few estimation approaches. Of the 16 studies we replicated, almost half (7) had at least one replication in which

[6] We calculated each p-value in the same way that the authors did, e.g. robust or clustered standard errors.

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 62 limited dependent variable models, the WTO variable was significant in 15 of them.

Replications with Existing Fixes

For the second set of replications, we incorporated different approaches that are designed to deal with unobserved heterogeneity or omitted variables. Some of the studies we replicated used these "fixes" in their published specifications, while others did not. Country fixed effects were the most commonly applied strategy for dealing with unobserved country-specific variation. 26 of the 94 replications used country fixed effects. Splines or some sort of time trend were the most commonly applied strategies for dealing with time-varying unobservables. 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 both.

To assess whether these "fixes fix," we began by stripping them out of all the replication specifications. We call these the "reduced" replications. The reduced replications are identical to the authors' original specifications in every way except (a) we added the WTO_{it} variable as in Equation 2 and (b) we did not include any fixed effects, splines, etc. We then applied each of these fixes one-by-one (and in combination with one another) to *all* replications. The result is that we can see 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 (e.g. splines). 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. If the original model had none of these, we added the common linear time trend. 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 IR research. The most common justification for the use of this technique is to address the issue of non-random selection or endogeneity. There are a plethora of matching techniques. Most are designed to better achieve balance on *observables*. While each matching procedure may differ in how it achieves balance on observables, each of them makes a common, and equally important, assumption about balance on *unobservables.* Since our concern is with the latter problem, we use the most commonly applied matching technique, propensity score matching.[8] Briefly, propensity score matching uses a set of observable variables to estimate the probability of a unit receiving treatment. Treatment here consists of joining the GATT/WTO. Treated and untreated observations with similar propensity scores are matched together, and then the dependent variable is compared across the treated and untreated observations.

For the matching replications, 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. The approach we use follows 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). In other words, what we have done here matches (no pun intended) standard practice in IR research. To be sure, there is much methodological debate over "what to match on." Some current practices emphasizes matching on observables which strongly predict treatment assignment. It is worth noting here that many of the observables in the studies we replicated fit this description. Many of the covariates used in the original replications to explain the dependent variable were also very likely associated with the propensity to join the GATT/WTO regime. For example, nearly all of the studies used things like GDP and democracy measures as covariates or control variables. These are strong predictors of whether and when a country is likely to join the WTO. It is also worth noting that the choice of which observables to match on is fundamentally a question of achieving balance on observables. Even when achieving perfect balance on observables, the problem of selection on unobservables potentially remains.

There are two important results from Table 1. The first is that the high rate of false positives is surprisingly persistent. When we simply added a WTO variable to the authors' regressions, we got a false positive rate of 34%. The false positive rate rises to 44% when we remove the authors' fixes. However, adding country fixed effects or country trends/splines only reduces the false positive rate back down to 34% for both. The matching approach fares similarly, with a false positive rate of

[7] The p-values are computed using the post-processed sample size.
[8] Rosenbaum and Rubin (1983).

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
CFE?	26/94				
Time?	72/94				
LDV?	62/94				

31%.

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

Recall, Figure 1 showed the p-values from the baseline replications, with the studies placed on the horizontal axis in ascending order of their p-values. 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 was significant in the original replications, but is insignificant with a commonly applied and relatively easy to implement fix.

However, there are also 8 blue O's in the bottom right quadrant. These are 8 new false positives resulting from the inclusion of country fixed effects. These O's denote studies for which the WTO variable was insignificant in the replication without country fixed effects, but is now significant with country fixed effects.

Figure 3 repeats the same process using the results from the matching replications. We retained the study order from Figure 1 and the vertical line denoting the break between significant and insignificant p-values. We see the same result as with Figure 2. There are 14 red X's in the top left quadrant- studies where the GATT/WTO variable was significant, but is insignificant when

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

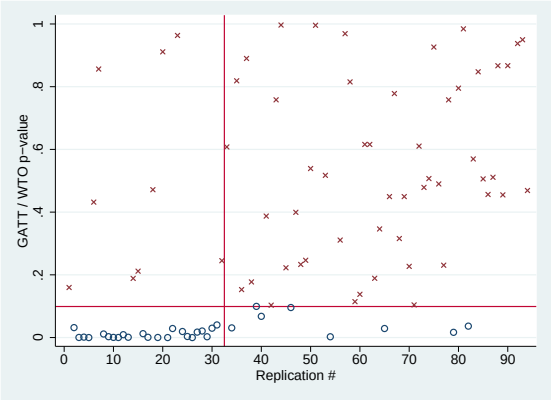
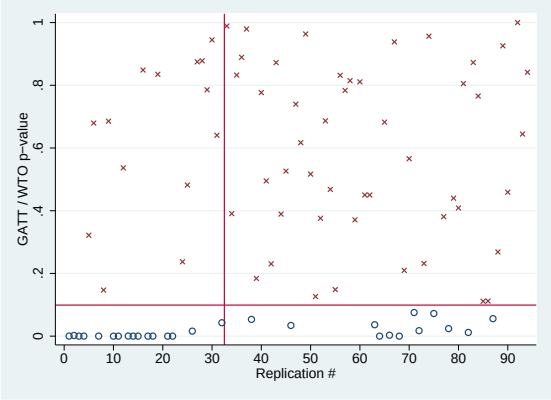


Figure 3: P-values for Effect of GATT/WTO on Irrelevent DV's, Matching



we use matching. However, there are 12 blue O's in the bottom right quadrant- new false positives that arise from the matching estimation 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 matching reduced bias or imbalance on observables between treated and untreated units varied across replications. However, achieving better balance on observables was not associated with an decreased probability of a false positive. The mean percent reduction in bias, averaged across each of the observables used in the replication, was 9.60% for replications that resulted in false positives, compared with 9.38% for replications that did not. A simple regression of the probability of a false positive on the percent reduction in bias shows virtually no association between the two.[9]

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

Combining Fixes

Table 1 showed that many of the commonly applied fixes do not lower the false positive rate substantially. 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 rate is almost twice the tolerable rate. 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. This would mean that some false positives could possibly be true positives. This is highly unlikely. To further provide evidence of this, we 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 facilitate "cooperation to safeguard certain species from over-exploitation."[10] CITES went into force in 1975 and 179 countries have become 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 we've analyzed. Its rules only govern the import and export of a minuscule percentage of global trade each year 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, and any positive results are attributable to selection on unobservables.

[10]CITES website, `http://www.cites.org/eng/disc/what.php`. Accessed 2-28-2014.

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

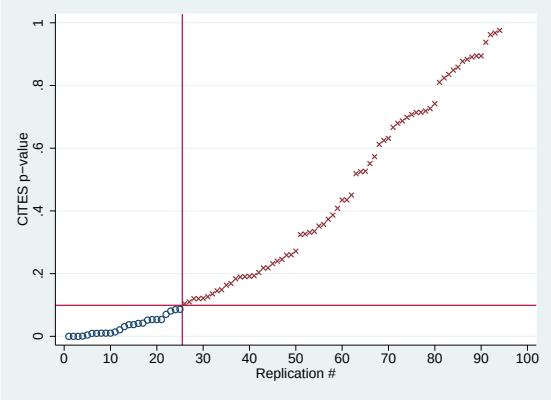


Table 3 replicates the results from the first table above. The false positive rates are only slightly lower than those found above. In the replications where we only added the CITES variable and left the authors' specifications otherwise unchanged, the false positive rate is 27%. In the reduced replications,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

As with the GATT/WTO replications, combinations of fixes also fail to lower the false positive rate with the CITES replications, as shown in Table 4, which repeats the same series of specifications as in Table 2. Our suspicion was that CITES false positives were driven strongly by

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

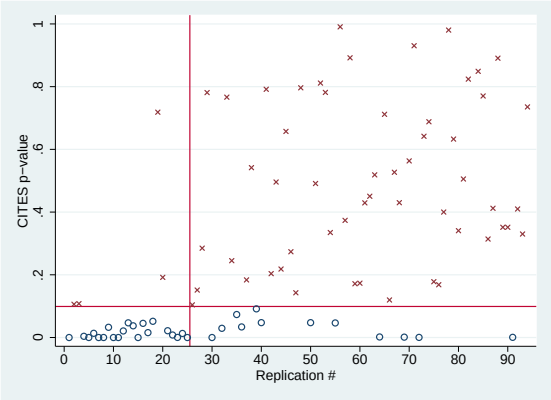
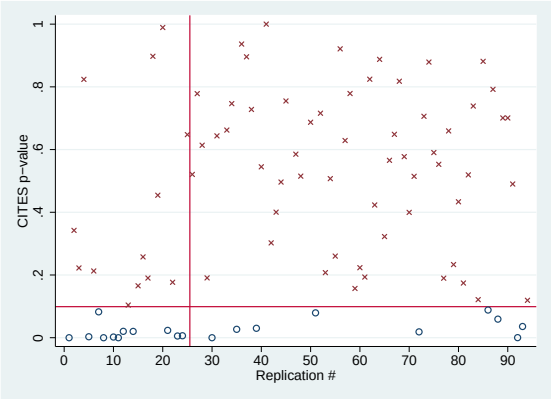


Figure 6: P-values for Effect of WTO on Irrelevent DV's, Matching



unobserved country-specific trends. This appears to be the case. 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.

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 severe problem and that existing fixes do not ameliorate this problem using real-world data and replications. This section generates intuition on why this is the case using a controlled environment where, unlike in the real world, 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 and the roles they play in confounding our ability to assess the effects of treaties on compliance. We describe each type mathematically and motivate them with ties to 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 constrains 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 recover too many 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 something akin to a "law of second best solutions." In economics, this term refers to situations where fixing one, but not all sources of 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 3 the ratification equation and Equation 4 the compliance equation.

$$r_{it} = 1(X_{it}\beta + u^r_{it} > 0) \tag{3}$$

$$c_{it} = X_{it}\beta + \alpha R_{it} + u^c_{it} \tag{4}$$

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 of the disturbances, we use the superscripts r and c to indicate whether the observable enters into the ratification or compliance equation.

Unobs. in ratification equation

$$u^r_{it} = \mu^r_i + \delta^r_t + \gamma^r_i t + e^r_{it}$$
$$\mu^r_i \sim N(m^r, \sigma^2_{r1})$$
$$\delta^r_t \sim N(d^r, \sigma^2_{r2})$$
$$\gamma^r_i \sim N(g^r, \sigma^2_{r3})$$
$$e^r_{it} \sim N(e^r, \sigma^2_{r4})$$

Unobs. in compliance equation

$$u^c_{it} = \mu^c_i + \delta^c_t + \gamma^c_i t + e^c_{it}$$
$$\mu^c_i \sim N(m^c, \sigma^2_{c1})$$
$$\delta^c_t \sim N(d^c, \sigma^2_{c2})$$
$$\gamma^c_{it} \sim N(g^c, \sigma^2_{c3})$$
$$e^c_{it} \sim N(e^c, \sigma^2_{c4})$$

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:

$$cov(\mu^r, \mu^c) = \rho_1$$
$$cov(\delta^r, \delta^c) = \rho_2$$
$$cov(\gamma^r, \gamma^c) = \rho_3$$
$$cov(e^r, e^c) = \rho_4$$

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 earliest 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, measure, 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 and measure precisely.

δ_t represents a year-specific component to the unobservables. This component describes factors which vary over time and which affect ratification and compliance. To continue the GATT/WTO and trade example from above, there are many possible candidates. Shipping costs are generally thought to have 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.

$\gamma_i t$ 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. Late joiners of the GATT/WTO may also have flatter trajectories of trade for the reverse reasons. 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, i.e. "this country was relatively more productive back in the day."

More complex types of unobservables, beyond the ones described above, 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 lots of 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

Given the high false positive rates in our replications and the theoretical complexity of the real-world DGP faced by IR scholars, we now show that our two main results from above, (1) that many fixes do not fix the problem of false positives and (2) fixes tend to make the problem better in some ways and worse in others, obtain even with simulations a simpler, known DGP.

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

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

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. Most applications treat compliance as binary: a member complies with the institution's rules or it does not. Compliance could also be continuous. For example, if compliance with the institution's rules required a country to lower its carbon emissions by x tons, then lowering emissions by $x > y > 0$ tons represents compliance to a certain degree.

The two simplifications for this DGP are as follows. First, we include only one covariate, x_{it}, which affects both membership and compliance. This is akin to assuming a DGP without

instruments which might only affect ratification or compliance but not both. Again, this affects our results and arguments quantitatively but not qualitatively, and these results are replicable with additional covariates.[11]

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:

$$u_{it}^r = \sqrt{.5}\mu_i^r + \sqrt{.5}\delta_t^r$$
$$u_{it}^c = \sqrt{.5}\mu_i^c + \sqrt{.5}\delta_t^c$$

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

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

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

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 experimental simulations, we set the number of units or countries to be $N = 100$ and the number of years to be $T = 30$. We chose these values because they were similar to those found in the observational studies that we replicated above. We set the variance of our observable

[11] We return to the topic of instruments in the conclusion.

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. The 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 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 that it can make performance worse when this is not the case. This second expectation is akin to a "law of second best." Depending on the nature of the DGP, using an approach that addresses one type of selection on unobservables can make results worse in the presence of other types.

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

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

	$\rho_\mu = 0$ $\rho_\delta = .5$ $\rho = .25$	$\rho_\mu = 0$ $\rho_\delta = .75$ $\rho = .375$	$\rho_\mu = .5$ $\rho_\delta = .5$ $\rho = .5$	$\rho_\mu = .75$ $\rho_\delta = .75$ $\rho = .75$
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 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. With fixed-effects, however, if only some of these sources can be eliminated, the estimator's performance can be worse than doing nothing. In other words, if there are three sources of endogeneity (e.g., unit effects, period effects, and unit-specific linear trends), the first-best solution is always to address all of them, but eliminating two of the three is not necessarily second-best. 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.

We find both of these trends in the results. Starting with continuous compliance Table 5, the performance of the do-nothing OLS and matching estimators gets progressively worse as we move from Case 1 to Case 4. It is worth noting that part of the problem for matching is the extreme overconfidence of the standard error estimates.[12]

[12]Following convention, we do not adjust our matching standard error estimates to account for the panel structure of our data.

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

	$\rho_\mu = 0$ $\rho_\delta = .5$ $\rho = .25$	$\rho_\mu = 0$ $\rho_\delta = .75$ $\rho = .375$	$\rho_\mu = .5$ $\rho_\delta = .5$ $\rho = .5$	$\rho_\mu = .75$ $\rho_\delta = .75$ $\rho = .75$
Logit				
Mean($\hat{\alpha}$)	.662	1.034	1.47	2.569
$\Delta \Pr(c_{it} = 1 \mid R_{it} = 1)$	.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
$\Delta \Pr(c_{it} = 1 \mid R_{it} = 1)$	.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%

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/membership and compliance, but does reduce significantly the variance in the unobserved disturbance that determines ratification/membership, it will make the simultaneity bias worse.

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 described later, for example, we are able to 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.

Turning to our binary outcomes, Table 6, we see mostly similar patterns across the cases, particularly if we compare OLS and the basic logit estimator. There are a couple of noteworthy differences, however. First, the false-positive rate of conditional logit is worse when compared to the linear fixed-effects estimator in Table 1. This reflects the poor performance of our bootstrapping technique. For panel models, the default bootstrap is typically to sample units with replacement. Since much of the dependence in our data is within-unit dependence driven by common period effects, it is likely that we could improve our standard error estimates by sampling cross-sections. Second, the false-positive rate of matching is better in the binary outcome case than in the continuous case. This is because the endogeneity problem stems from the covariance between the latent compliance and ratification/membership propensities. Since the logit models are parametric models of latent compliance, these estimators are more vulnerable to the simultaneity bias. If we think of the binary compliance outcome as a very rough proxy for latent compliance, the measurement error in binary compliance fortuitously attenuates the simultaneity bias for the matching estimator. A more positive way to look at it, perhaps, is that the non-parametric nature of the matching estimator protects it slightly against the simultaneity bias.

Sensitivity Tests

How can researchers assess the likelihood that their results are "true positives" compared to "false positives"? A wide array of sensitivity tests are designed to help with this question. In general, they ask how sensitive particular estimates are to the presence of unobservables which influence selection and outcome. Our purpose here is twofold. First, we delineate the components which are common to the vast majority of sensitivity tests. We use these components to classify two types of tests – one which benchmarks results against the selection process and one which benchmarks against the outcome process – using some of the replications above. To the best of our knowledge, this is the first political science application of the former.

A second goal of this section is to give practical guidance on the conditions under which each type of test is appropriate. An under-emphasized feature of sensitivity tests is that, just like any one proposed estimation approach, no one sensitivity test can be a panacea for selection problems. They each require the researcher to draw on her prior knowledge to benchmark the results, and the implications of each test depends on the confidence in that knowledge. The strength of the researcher's beliefs in her prior knowledge affects the strength with which she can draw conclusions from each sensitivity test.

Parts of Sensitivity Tests

A defining feature of many sensitivity tests is that they focus on the extent to which selection on unobservables influences the estimated treatment effect. Broadly speaking, there are two approaches: (1) make an assumption about the unobservables and estimate the treatment effect and (2) make an assumption about the treatment effect and evaluate its implication for the unobservables.

With the first approach, the researcher evaluates the null hypothesis of no treatment effect under an assumption that is more conservative than the standard assumption that treatment is conditionally independent of relevant unobservables. This is typically done by estimating a constrained selection model, where the constraint imposes the new, more conservative assumption about selection on unobservables. If the estimated treatment effect maintains its sign and statistical significance, we conclude that the original estimated effect is robust, and likely not a false positive. One advantage of this approach is that one does not need instruments to estimate the treatment effect. The drawback is that estimation under the constraint can be difficult.[13]

The second approach is generally less formal, but easier to implement. With this approach, one estimates the model under the assumption that the null hypothesis of no treatment effect is true and then calculates the *implied* difference in unobservables that would produce a given treatment effect estimate. In other words, the second approach reverses the first. The researcher makes an assumption about the treatment effect and evaluates its implication for unobservables. If the implied features of the unobservables is unreasonable, we conclude that the estimated effect is robust to selection on unobservables.

Sensitivity tests have three parts: alternative assumption(s), a quantity of interest, and a benchmark. The two tests that we discuss have similar alternative assumptions. They each assume that the true treatment effect is different from the one originally estimated (e.g. the true treatment effect is zero) and assume the presence of selection on unobservables. The quantity of interest is the implied features in unobservables that follows from this assumption. The two specific approaches we use focus on different implied features of the unobservables. One focuses on implied differences in unobservables across treatment and control units. The other focuses on implied explanatory power of unobservables.

The third component, the benchmark, is the most important, but also the least emphasized in

[13]Both Altonji et al. and Imbens estimate constrained selection models. Altonji et al. experienced significant convergence problems and had to eliminate a number of covariates to estimate their constrained bivariate probits, (fn. 15, p. 166). The point is that estimating these models is not always a simple exercise, especially when the set of covariates is large.

methodological work describing sensitivity tests. The crucial question is: what is a "reasonable" or "unreasonable" implied feature of the unobservables? The researcher must gauge the plausibility of these implied features against something she knows about the real world. This can either be what she knows about the selection (i.e., ratification) process or the outcome (i.e., compliance) process. The statistical results from any sensitivity test have little epistemological traction *until they are benchmarked relative to the researcher's beliefs*. Which sensitivity test is most appropriate thus depends crucially on what the researcher thinks she knows about the process against which she will ultimately benchmark her results. Her beliefs are formed from existing knowledge, related work, or logical argument. If she feels confident about her understanding of the selection process, then she should use the first sensitivity test we present, and if she feels confident in her understanding of the outcome, the second.

Benchmark 1: Selection on Observables

The first specific approach we consider is from Altonji, Elder and Taber (2005). With this approach, the alternative assumption is that the null hypothesis of no treatment effect is true. The quantity of interest is a ratio: the amount of selection on unobservables relative to selection on observables that is implied by this assumption. In other words, this approach uses "the degree of selection on observables as a guide to the degree of selection on unobservables (Altonji, Elder and Taber, 2005, p. 153)." The approach 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 the treatment on the outcome? If this ratio is high, *relative to what the researcher knows about the selection process*, then the researcher is more confident in her original result. The phrase in italics highlights the benchmark being applied: if the researcher is confident in her knowledge about the selection process and its strength, then this ratio can be a powerful tool to assess results. If she is not confident, then this ratio is less informative.

To be concrete, consider our compliance regression equation,

$$c_{it} = \alpha r_{it} + X'\gamma + u^c_{it}.$$

In this linear case, if we assume $\text{cov}(r_{it}, X) = 0$, the ordinary-least-squares (OLS) estimate for the

treatment effect is

$$\begin{aligned}\text{plim } \hat{\alpha} &= \alpha + \frac{\text{cov}(r_{it}, u^c_{it})}{\text{var}(r_{it})} \\ &= \alpha + E[u^c_{it}|r_{it} = 1] - E[u^c_{it}|r_{it} = 0],\end{aligned}$$

where $E[u^c_{it}|r_{it} = 1] - E[u^c_{it}|r_{it} = 0]$, the difference in the expected disturbance (i.e., the unobservable determinant of compliance) across the treatment and control groups, is the simultaneity bias in $\hat{\alpha}$. In other words, the bias is a function of the imbalance in unobservables across the treatment and control groups. In order to make the treatment and covariates orthogonal, Altonji et al. suggest estimating

$$r_{it} = X'\beta + u^r_{it}$$

and substituting the ratification disturbances into the compliance equation, which gives

$$c_{it} = \alpha u^r_{it} + X'(\gamma + \alpha\beta) + u^c_{it}.$$

If X is orthogonal to u^c_{it}, then $\text{cov}(r_{it}, u^c_{it}) = \text{cov}(u^r_{it}, u^c_{it})$, but $\text{var}(u^r_{it}) < \text{var}(r_{it})$. Hence, the numerator in the simultaneity bias must be adjusted, giving

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

From this we can calculate the quantity of interest

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

which is the imbalance in unobservables under the null hypothesis, $\alpha = 0$.

In order to know whether such an imbalance is plausible or not, we need a benchmark. Altonji et al. recommend benchmarking against the imbalance in observables. More formally, they note that standardized selection on unobservables is equal to standardized selection on observables when:

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

Thus, we can calculate $\hat{\alpha}\frac{\text{var}(u^r_{it})}{\text{var}(r_{it})\text{var}(u^c_{it})}$, the standardized implied imbalance in unobservables under the null hypothesis $\alpha = 0$, and compare it against our benchmark, the standardized im-

balance on observables. If $\hat{\alpha}\frac{\text{var}(u_{it}^r)}{\text{var}(r_{it})\text{var}(u_{it}^c)} > [E[X'\gamma|r_{it}=1] - E[X'\gamma|r_{it}=0]]/\text{var}(X'\gamma)$, selection on unobservables is stronger than selection on observables under the null hypothesis. When the left-hand side of the inequality is substantially larger than the right-hand side, we conclude that the necessary imbalance is implausibly large and the estimated causal effect is robust to selection on unobservables. If, on the other hand, $\hat{\alpha}\frac{\text{var}(u_{it}^r)}{\text{var}(r_{it})\text{var}(u_{it}^c)} < [E[X'\gamma|r_{it}=1] - E[X'\gamma|r_{it}=0]]/\text{var}(X'\gamma)$, the implied imbalance in unobservables under the null hypothesis is less than the imbalance on observables. In this case, we typically conclude that the estimated causal effect is *not* robust to selection on unobservables.

Benchmark 2: Explanatory Power of Observables

Our second specific approach uses the explanatory power of observable covariates as a benchmark (Imbens, 2003; Blackwell, 2014). This approach is better known in political science, so our description will be more brief. In our compliance framework, the point-biserial correlation coefficient between the binary ratification variable and the continuous unobservable determinant of compliance is

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

where n is the total number of observations, n_1 is the number of observations under treatment, and n_0 is the number of observations under control. From this we can calculate the proportion of the variance in outcomes that would be explained by the implied imbalance in unobservables across the treatment and control groups under the null hypothesis, $\alpha = 0$:

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

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 the most powerful covariate, we conclude that such an imbalance is unlikely and that the causal effect estimate is robust to selection on unobservables. If, on the other hand, 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 estimated causal effect is sensitive to selection on unobservables.

Which Benchmark?

Our advice regarding the appropriate benchmark is simple. When the researcher knows a lot about the selection 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 selection process or knows more about the outcome process it is better to benchmark against the explanatory power of covariates, their ability to explain the outcome, regardless of whether they are balanced across the treatment and control groups (the second approach).

Knowledge of the selection process consists of knowledge of the set of variables which do and do not affect selection into treatment. The Altonji approach uses selection on observables as a guide for selection on unobservables. Without knowledge of the selection process it is impossible to know whether this is a reasonable guide or not.

Consider the numerator and denominator in the standardized selection on observables calculation, the left hand side of Equation 5. One could just *assume* everything that determines the outcome also drives selection, but this may or may not be the case, and if it is not, the standardized difference in observables is not a useful benchmark in that it will not effectively screen false positives. Variables that are irrelevant to (or independent of) the selection process will be balanced across the treatment and control groups, but as long as they help to explain the outcome, these covariates will increase the variance of the linear predictor $X'\gamma$. Thus, whenever the linear prediction contains irrelevant variables, it will shrink the selection on observables calculation toward zero.

The relevant variables are the set of covariates that determine both selection into the treatment (ratification) and the outcome (compliance). It is important to be selective when choosing this set because including irrelevant variables will reduce the power of the test to detect false positives. Knowledge of the selection process is important in another respect as well. The researcher needs to be confident that the direction of the imbalance in unobservables, the sign of $E[u^c_{it}|r_{it} = 1] - E[u^c_{it}|r_{it} = 0]$, is the same as the imbalance in the relevant observables, the sign of $E[X'\gamma|r_{it} = 1] - E[X'\gamma|r_{it} = 0]$, and this requires some knowledge of the likely unobservable forces at work. Obviously, if there is reason to believe the sign of the difference in observables and the sign of the difference in unobservables are not the same, one should not use selection on observables as a guide to selection on unobservables.

Finally, whenever the researcher knows little or even nothing about the selection process, it is better to benchmark against the explanatory power of covariates, their ability to explain the out-

come, regardless of whether they are balanced across the treatment and control groups. However, this approach faces a similar problem when the researcher lacks knowledge of strong predictors of the outcome. If the researcher is not confident in the explanatory power of particular observables, then finding that unobservables would need to be stronger than certain observables does not help the researcher eliminate false positives.

Two Illustrations

To illustrate these two choices 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 tax revenues and lower illiteracy rates. We concentrate on these two outcomes.

In our replication analyses, we found a positive relationship between GATT/WTO membership and tax revenues and we also found a positive relationship between CITES membership and illiteracy rates. We – the researchers – strongly suspect that both results are false positives. It is highly unlikely that WTO membership increases tax revenues, especially since the WTO limits the use of revenue-raising tariffs. It is also highly unlikely that the CITES treaty has any true effect on literacy rates.

We also feel more confident in our knowledge regarding the selection process for the GATT/WTO than we do for the CITES treaty. Recent work has (Davis and Wilf, 2011) has linked WTO membership with political and economic indicators that are present in the Gerring et al. studies. We know much less about what causes countries to join CITES, and we doubt that many of the variables in the Gerring et al. study are strong predictors of CITES membership.

In this section, we illustrate how we can leverage our knowledge of GATT/WTO membership to assess the (likely) false positive relationship between GATT/WTO membership and tax revenues. Because of this knowledge, we can use the first sensitivity approach to rule out the positive relationship between WTO membership and tax revenues as a false positive.

We then demonstrate how our limited knowledge of selection into the CITES treaty limits our ability to use this same approach to assess the false positive relationship between CITES and literacy. We don't know enough about selection into the CITES treaty to rule out this false positive.

However, we do have stronger knowledge of the explanators of literacy rates (e.g. the explanators thoughtfully chosen in the Gerring et al. study itself). This allows us to use the second approach and benchmark our sensitivity tests against the strength of observables in explaining outcomes, which allows us to assess this false positive more confidently.

Tax Revenue False Positive

We start with the tax revenue regression to which we added the GATT/WTO treatment variable. Our estimated treatment effect suggests that GATT/WTO membership increases a country's tax revenue as a share of GDP by 3.69%. The estimated coefficient is statistically significant with an associated t-statistic is 6.96. The result is robust to including fixed country effects, and we find a statistically significant positive treatment effect when we match on the covariates. This positive relationship is almost certainly spurious. If there is any direct causal effect of GATT/WTO membership on tax revenue, it should be negative since membership requires countries to lower their tariffs, but does not require other changes to their tax policy. The tax revenue data used by Gerring et al counts tariffs as tax revenue.

This is a case where we know something about the selection process. We draw on new work by Davis and Wilf (2011) to identify the covariates in Gerring et al.'s tax regression that also explain GATT/WTO membership. Davis and Wilf analyze the economic and political bases of membership. Countries join for the economic benefits, but existing members also control the accession process to achieve geopolitical objectives. Two sets of variables in the Gerring et al regression are relevant. The first includes covariates that describe a country's regime type. Davis and Wilf describe GATT/WTO as a "like-minded club" that admits new members who share important characteristics with existing members, and they find in their empirical analysis that a country's Polity (democracy) score robustly predicts its time to GATT/WTO application. The second set includes covariates that relate to a country's economic size. A number of theoretical arguments emphasize economic size as a determinant of GATT/WTO membership. For instance, large-economy countries benefit from making commitments that prevent them from using tariffs to extract rents. Empirically, Davis and Wilf find that both GDP and GDP per capita explain time to application. Based on our knowledge of the political and economic bases of membership, 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, and this may discourage membership among oil exporters.

The standardized implied imbalance in unobservables under the null hypothesis is .08,

$$\hat{\alpha}\frac{\text{var}(u_{it}^r)}{\text{var}(r_{it})\text{var}(u_{it}^c)} = 3.69\frac{.11}{(.17)(30.49)} = .08.$$

Using the five covariates that are also relevant to GATT/WTO membership, the standardized imbalance in observables is .18,

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

Thus, under the null hypothesis of no treatment effect, selection on unobservables would only have to be .44 as strong as selection on the relevant observables. This seems plausible, which suggests it is likely that we have a false positive.

The sensitivity test proposed by Altonji et al. is an effective screen in this example because we know something about the selection process. But this need not be the case. If we were to apply the same calculations using *all* the covariates included in the Gerring et al. tax regression, we would come to a different conclusion. The standardized difference in observables across the treatment and control groups when we include all the covariates is .05,

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

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 seems far less likely than .44 as strong and suggests the estimated treatment effect may be a true positive. Notice that the calculations differ mainly because of the standardization. The numerators are fairly similar in size, but the denominators are quite different.

The linear prediction $X'\gamma$ when we include all of the covariates has a lot of noise in it that is orthogonal to selection. These irrelevant variables are balanced across the treatment and control groups, and so the numerator is the standardized difference in observables is about the same as when we calculate using only the variables that are relevant to both selection and the outcome, but the denominator is much larger.

Literacy False Positive

Next, we turn to the Gerring et al. illiteracy regression. When we include our CITES variable, we find that ratifying the treaty causes the illiteracy rate among adults to rise by approximately

16%. For example, at the average values for the other covariates in the model, the illiteracy rate increases from 5.00% to 5.87% points. The coefficient estimate (.16) is statistically significant with a t-statistic of 2.57. We are nearly certain that this is a false positive; we are unaware of any relationship between an environmental treaty and literacy rates. Yet the result is robust to including fixed country effects, and matching also generates a statistically significant treatment effect.

Can sensitivity analysis help us? In this case, selection on observables is not a useful guide for benchmarking selection on unobservables. The standardized difference in observables across CITES and non-CITES countries is negative, meaning that we expect, on the basis of observable differences, lower illiteracy in countries that have ratified CITES. If we believed that selection on unobservables were equal to selection on observables, we would conclude that the simultaneity bias is attenuating, that we are underestimating the causal effect of ratifying CITES on the illiteracy rate. This is even more implausible than the original false positive.

We could try to pare down the list of covariates, as we did previously with the GATT/WTO sensitivity analysis, but we do not have a good theory regarding the selection process. We do not have other work or prior knowledge that we can draw on to gain leverage over the causes of CITES membership. Thus, this example demonstrates why it is not a good idea to benchmark on selection when one does not know the relevant set of covariates.

Fortunately, we can benchmark against the explanatory power of the covariates in the Gerring et al. regression instead. The partial coefficients of determination for the covariates in their model range from 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,

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

This is lower than all but 3 of the partial coefficients of determination for the variables in Gerring et al.'s model. Thus, using this benchmark, we can see that 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 more 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 the first step to recovery. There are strong theoretical reasons to expect that unobservables affect ratification and compliance. This phenomenon generates false positives, where we mistakenly conclude that certain institutions cause compliance. As we show with a replication exercise based on existing work and with Monte Carlo simulations, this problem is likely to be severe and multifaceted. We found false positive rates which were generally around 34%, which is much higher than would be tolerated by conventional assessments of statistical significance.

Second, there is not 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 estimate her quantities of interest under more conservative assumptions than those generally used. The researcher benchmarks the conservativeness of these assumptions according to her theoretical knowledge. We presented two sensitivity approaches here. One, which has been previously unused in political science, gives the researcher power based on her knowledge of the selection process. The other, more commonly used approach, gives the researcher power based on her knowledge of the explanatory power of certain covariates. Ultimately, the persuasiveness of these approaches is founded on the researcher's theoretical knowledge against which she will benchmark her results.

Fourth, this paper has not discussed instrumental variables approaches. Researchers rarely have access to a valid instrument, i.e. something measurable which is correlated with treaty membership but not compliance. We leave it to particular applications to assess whether these assumptions are plausible or not. *ex ante*, and without theoretical knowledge of a particular context or application, we are hesitant about claims that a certain variable affects ratification but not compliance. Both

decisions, ratification and compliance, are made by similar (often identical) political actors, who have similar interests and informational endowments about the costs and benefits of their decisions. Even when the actors making these two decisions are distinct, it is highly unlikely that those actors are making these decisions in isolation of one another. Factors, be they observable or unobservable, which affect compliance change the leader's calculus over whether to ratify and vice versa. Often, the theoretical argument used to motivate the relationship between the instrument and ratification (i.e. an argument for the instrument's strength, which fortunately, is testable) applies equally well to a possible relationship between the instrument and compliance (i.e. an argument about exogeneity, which unfortunately, is not testable). While instrumental variables are a very appealing approach for assessing ratification and compliance, they may prove difficult in practice.

Finally, the modal phrase in this conclusion, by far, has been "theoretically informed." 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. At the end of the day, there is no substitute or alternative to theoretically informed research.

References

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.

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

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

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.

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.

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

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

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.

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