Stephen Chaudoin
Department of Political Science
University of Illinois at Urbana-Champaign
chaudoin illinois.edu

Jude Hays
Department of Political Science
University of Pittsburgh
jch61 pitt.edu

Raymond Hicks
Department of Politics
Princeton University
rhicks princeton.edu
*

2   CHAUDOIN, HAYS, HICKS

We use a replication experiment of 94 specifications from 16 different studies to show the severity of the problem of selection on unobservables. Using a variety of approaches, we show that membership in the GATT/WTO has a significant effect on a surprisingly high number of dependent variables (34%), variables which have little to no theoretical relationship to the WTO. To make our exercise even more conservative, we demonstrate that membership in a low-impact environmental treaty, CITES, yields similarly high false positive rates. We advocate for theoretically informed sensitivity analysis, showing how prior theoretical knowledge conditions the crucial choice of covariates for sensitivity tests. While we focus on international institutions, the arguments apply to other subfields and applications.

A fundamental question in International Relations research asks whether ratifying or joining an international institution affects the policies of sovereign nation states. Research in this vein 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 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, & Barsoom, 1996). Even if ratification has no 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 relationship, or if is only an artifact of selection on unobservables.

Researchers outside of International Relations (IR) face similar challenges. 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.

This paper seeks to make two contributions. The first is serious assessment of the severity of the problem of selection on unobservables. Extant research, in IR and beyond, 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 demonstrates a relationship between membership and compliance, as opposed to a false positive?

We present evidence from a novel, 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 inference. The results are also of substantive significance, with GATT/WTO membership having a meaningful effect on these dependent variables. 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.

4   CHAUDOIN, HAYS, HICKS

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 Endangered Species (CITES). It is very unlikely that CITES, which institutes licensing requirements for a small amount of endangered plants and animals, has any relationship with the dependent variables in our replication exercise, none of which describe environmental outcomes. Yet, we again find high false positive rates. This gives convincing evidence that our findings are not merely the result of true relationships which researchers do not yet understand.

In addition to demonstrating a very high false positive rate, the replication exercises also demonstrate a subtler pattern. 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. Addressing one type of unobservables, without addressing all types of unobservables, can often make the problem of false positives worse. This phenomenon is a type of "law of second best," where addressing one type of unobservables can be worse than addressing none.

The second contribution of this paper is to advocate for theoretically informed sensitivity testing. Sensitivity analysis is a powerful tool for assessing the likelihood that a positive result is a false positive. However, the leverage of a particular sensitivity test depends on the theoretical knowledge against which the test is benchmarked, not just the mechanics of the approach. We demonstrate this using a sensitivity approach based on Altonji, Elder, and Taber[1]. This approach asks "How severe would selection on unobservables need to be, relative to selection on observables, to account for the estimated effect of ratification on compliance?"

Our contribution to this approach is to show how prior theoretical knowledge is crucial to the choice of which covariates to include in the sensitivity test, which in turn, has significant effects on the ability of the test to screen false positives and retain true positives. For applied research, the choice of covariates for a sensitivity test is an equally important concern to the mechanics of a particular sensitivity approach. We use examples from our replication exercise to demonstrate how the approach can succeed or fail, depending on the strength of this theoretical knowledge regarding the covariates selected for the sensitivity test. These examples also give practical advice for applied researchers on how to use the approach and assess its strength. In addition, we provide an original Stata command for the general implementation of these approaches, which we will make publicly available. Our goal is to make this type of testing more widespread and accessible, while still retaining a transparent, concrete emphasis on the theoretical knowledge underlying the results.

Lastly, 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 the context of international institutions and compliance. 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.

---

[1] Altonji et al. (2005)

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. The dependent variable often describes 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[2], Simmons and Hopkins[3] and Von Stein[4] 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. There are many examples of similar phenomena outside of IR, where unobservables make selection into a particular treatment or regime non-random, which potentially biases the resulting estimates of the effect of treatment on outcome.

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.[5] The vector X_{it} contains the observable characteristics of a country which potentially affect compliance and ratification. u_{it}^r and u_{it}^c are unobservables that affect ratification and compliance respectively.[6]

$$r_{it} = f(X_{it}B + u_{it}^r) \qquad \text{(Ratification Equation)}$$
$$c_{it} = f(X_{it}\beta + \alpha r_{it} + u_{it}^c) \qquad \text{(Compliance Equation)}$$

$$(1)$$

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 biases estimates upwards. As a consequence, even when we find positive estimates of α,

---

[2] Simmons (2000)

[3] Simmons and Hopkins (2005)

[4] Von Stein (2005)

[5] Compliance need not be binary. In the appendix, we consider both continuous and binary measurements of compliance.

[6] 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.

6   CHAUDOIN, HAYS, HICKS

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.[7] While it is theoretically possible to look for sources of exogenous variation in treaty membership, as in an instrumental variables or natural experiment approach, such sources are highly unlikely to exist given that largely the same actors make both the ratification and compliance decisions.

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. [8]

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 a later 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. The studies we replicate were generally not related to treaty ratification, so if we find high false positive rates in our replications, we should be concerned that false positive rates may be even higher in studies of ratification, where the selection on unobservables problem is potentially more severe.

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.

It is also important to note that we find many examples of *substantively* meaningful effects among these placebo tests. We discuss several of these examples in the sensitivity section. Other examples of substantively important findings include: GATT/WTO membership increases the probability a country has a legislature by 6%, decreases the presence of governmental torture by 4.3%, and increases life expectancy by $\sim 2\%$, among others. CITES membership also had

---

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

[8] A growing body of literature also discusses the reliability of treatment effects estimates. Angrist and Krueger[9], for example, discuss the strengths and weaknesses of different identification strategies such as OLS, fixed effects, instrumental variables, and matching. Rosenbaum[10] focuses on the use of sensitivity analysis as a way to more accurately estimate treatment effects. Both are important and we extend their advice by examining whether different identification strategies do solve the false positive issue and providing guidance on sensitivity testing.

substantively important estimated effects, such as decreasing infant mortality rates by 8.1% and decreasing the probability of political instability by 43%, among others. We focus on statistical significance to compare across replications, but our estimates also indicate suspiciously strong substantive relationships between membership in GATT/WTO/CITES and theoretically distant dependent variables.

*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.[11]   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.[12] 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. The studies varied in how they justified their empirical approaches, with some explicit about the assumptions underlying their chosen model and others less so. The studies also varied in the degree to which they argued that their approach was likely to be susceptible to the issue of selection on unobservables.

*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.[13] 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 line marks the 0.10 level. The vertical line marks the 32nd replication, which is the replication with the greatest p-value that still falls below the 0.10 threshold.

The two lines divide the figure into four quadrants. 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. O's in the bottom left correspond to "false positives," 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. 31 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.

---

[11] We had to limit ourselves to studies where the authors provided replication materials online or upon request. We added one study from *ISQ* which used country-year units of observation and which also devoted significant attention to the problem of selection on unobservables. A full list of the studies is available in the appendix.

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

[13] 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[14].

8   CHAUDOIN, HAYS, HICKS

Fig. 1. P-values for Effect of WTO on Irrelevant DV's
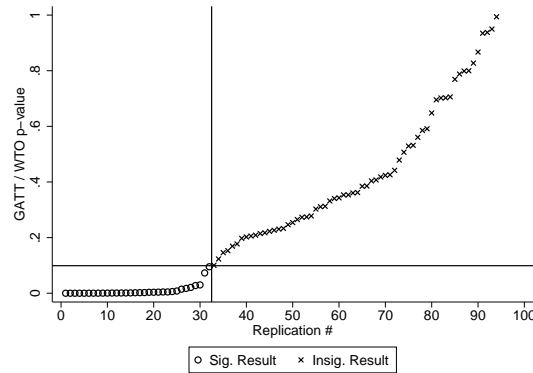
The false positives are also not concentrated in just a few studies or estimation approaches. Of the 16 studies we replicated, almost half (7) had at least one replication in which the WTO variable was statistically significant. Of the 34 different dependent variables analyzed in the 16 studies, the 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. The false positives are also not strongly correlated with the subject matter of the replication study or the number of countries or years in its sample.[15]

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. For example, researchers often argue in favor of using unit fixed effects to control for unobservables.[16] 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.

---

[15] See appendix for more details. We thank an anonymous reviewer for this suggestion.

[16] Keele[17] notes that this is an appropriate strategy for identifying causal effects if the researcher believes that unobservables are unit-specific and time-invariant.

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.[18] Matching techniques, in which the sample is pre-processed or pruned, are often used. In applied research, a very common justification for using this technique is to address non-random selection or endogeneity.[19] We use one of the most common matching techniques, propensity score matching.[21] 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 the advice of Ho, Imai, King, and Stuart[23]: "All variables in X_i that would have been included in a parametric model without preprocessing should be included in the matching procedure" (216).[24] 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.[25]

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 in the matched sample, it is still possible for inferences to be biased because of imbalance on unobservables (Sekhon, 2009). For example, in simulations presented in the appendix, we can achieve very good balance on observables 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 stan-

---

[18] The p-values are computed using the post-processed sample size.
[19] Miller[20].
[21] Rosenbaum and Rubin[22].
[23] Ho et al. (2007)
[24] 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.
[25] We used *psmatch2* in Stata, Leuven and Siansei[26].

10   CHAUDOIN, HAYS, HICKS

dard, 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%	29%	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 rates to 29% and 34%, respectively. The matching approach fares similarly, with a false positive rate of 31%.

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 lines. For Figure 2, X's still denote insignificant p-values, greater than 0.10, and 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 14 X's in the upper left quadrant of the figure, which denote the 14 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 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 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 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

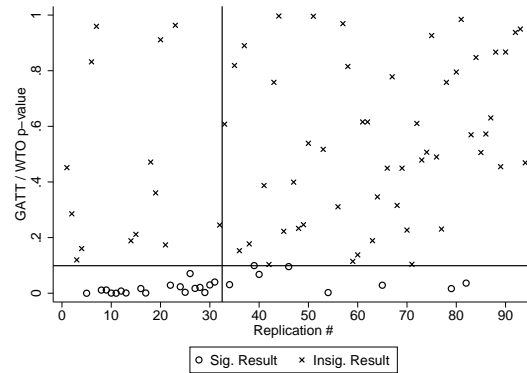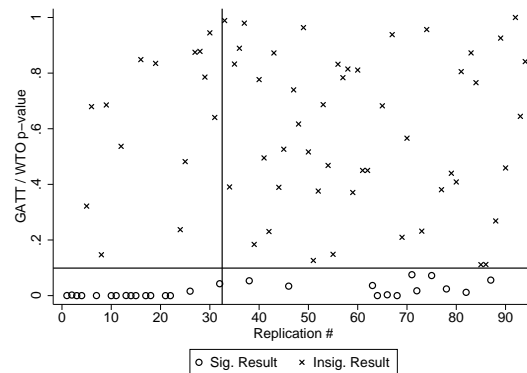*Fig. 2. P-values for Effect of GATT/WTO on Irrelevant DV's, Fixed Effects*

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

percent reduction in bias shows virtually no association between the two.[27] And to reiterate, in the simulations contained in the appendix, 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 in the matched sample.

Combining Fixes So far we have only referred to unobservables *writ large* and assessed whether individual approaches decreased false positive rates. Yet, unobservables come in many types. Some are country-specific and time-invariant. 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

---

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

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.

Some unobservables may 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.

Some unobservables may take the form 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.

Given that there are many possible types of unobservables present, do combinations of fixes, with different fixes designed to address different types of unobservables, lower the false positive rate? Here, we show 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, relative to fixing none of the market imperfections. A similar phenomenon occurs here. 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.

The first-best solution is to use an empirical approach that eliminates *all* of the unobservables that generate spurious sources of covariance between ratification and compliance. If this can be done, the effect of ratification on compliance 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. In related work, Plumper and Troeger[28] finding that unit-fixed effects strategies may be worse than pooled strategies in the presence of unobserved trending. Clarke[29] and Clarke[30] yield a similar finding, that inclusion of control variables has complex, possibly undesirable, effects on bias. Including an additional control variable could increase or decrease bias in the resulting estimates of interest.[31]

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

[28] Plumper and Troeger (2013)
[29] Clarke (2005)
[30] Clarke (2009)
[31] For more general discussions of a similar phenomenon, see Pearl[32] or Spirtes, Glymour, and Scheines[33].

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%	20%	36%	20%
No. Replications	88	91	91	93
No. Studies	16	16	16	16

One example of the law of second best comes from examining false positive rates in the original replications, the replications with country trends, and the replications with country trends and country fixed effects. The original false positive rate was 34%, and it decreases when adding country trends. A researcher might reasonably expect that there are country-specific, time-invariant unobservables that she might want to address. However, adding country fixed effects to the country trends raises the false positive rate to 20%.

CITES One possible concern is that the GATT/WTO regime truly does have an 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 effect on the dependent variables we analyze. The replications with CITES also have the advantage that, unlike the GATT regime, it is not simply developed Western democracies who join the regime early on. CITES members are a diverse group, and the earliest members included countries with the most endangered species in need of protection.

Table 3 replicates the results from the first table above. The false positive rate, 27%, is 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 false positive rate to 27% and 22% respectively.

The same problem found above, where fixes remove some false positives while also creating new ones, is again present. Figures 4-6 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. Figures 5 and 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

14   CHAUDOIN, HAYS, HICKS

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

Fig. 4. P-values for Effect of CITES on Irrelevant DV's
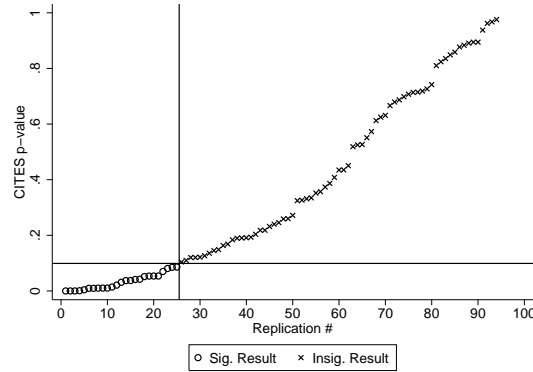
insignificant before. Matching fares slightly better, removing 13 false positives, but creating 9 new ones.

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 country and year fixed effects, but is still too high (25%). Year fixed effects in isolation yield a very high false positive rate, 42%. Adding country fixed effects to country trends again raises the false positive rates from 26% to 31%.

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	26%	31%	42%	25%
No. Replications	88	92	90	93
No. Studies	16	16	16	16

Simulations We have focused on our replications because they provide tangible, real-world examples of the situations and decisions facing applied researchers. However, all of these results are replicable in a controlled environment using Monte Carlo simulations. The appendix contains an extensive description of these simulation results. We describe a general data-generating process (DGP) that is theoretically grounded in our understanding of treaties and compliance

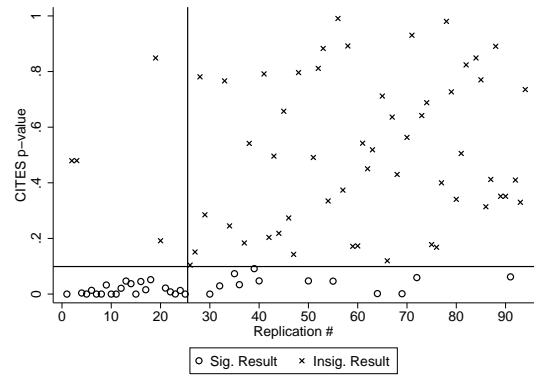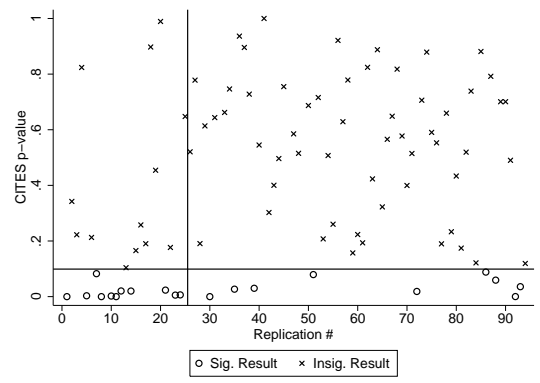*Fig. 5. P-values for Effect of WTO on Irrelevant DV's, Fixed Effects*

Fig. 6. P-values for Effect of WTO on Irrelevent DV's, Matching

16   CHAUDOIN, HAYS, HICKS

and which accommodates several possible types of unobservables. We then consider results from four cases of replications. The cases differ from one another in two ways. First, we gradually increase the overall covariance between the ratification disturbance term and the compliance disturbance term. In other words, the overall problem of selection on unobservables gradually gets worse.

Second, we vary the type of correlation across disturbances. In some cases, all of the covariance between ratification and compliance disturbances is attributable to within-unit variance caused by our period effects. In other cases, this covariance is attributable to both within- and between-unit variance in the unobservables. In other words, some cases involve only one type of selection on unobservables, and others involve two sources.

We evaluated the performance of three approaches: OLS without any fixed effects ("do nothing"), unit fixed-effects, and matching. We expected and found two trends in the results. First, the false-positive performance of the "do-nothing" estimators deteriorates 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 improved in our high covariance cases where some of the overall covariance is attributable to unit effects, but deteriorated when this is not the case.

Additionally, the false positive rates of the matching approach 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 false positives. 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

Unobservables affecting ratification and outcomes like compliance are likely to be complex and multifaceted. Applied empirical work risks producing biased estimates when assumptions about unobservables do not match the "true" DGP. This is particularly daunting since assumptions about unobservables are inherently untestable.

Sensitivity analysis is a powerful tool for conditioning inference even when the true nature of unobservables is unknown. We advocate for sensitivity analysis that uses observables as a guide for assessing the consequences of unobservables (Altonji et al., 2005). This type of sensitivity analysis asks "how severe would selection on unobservables need to be, *relative to selection on observables*, to drive our estimated effect to zero?" The approach compares the marginal effects of theoretically-relevant, measurable covariates – observables – and unobservables on the probability of ratification, i.e. of receiving the treatment. If the conclusion is that selection on unobservables would need to be twice as severe as selection on observables, for example, this means that the marginal effect of unobservables on the probability of receiving the treatment would have to be twice as large as the marginal effect of observables.

In practice, this approach requires the researcher to choose the observable covariates which will serve as the reference set for benchmarking the strength of unobservables. This set can include any number of the observable covariates used in the analysis. Observables in this context, $X'\beta$, are a linear combination of covariates with weights that reflect their marginal effect on the outcome. We demonstrate how the choice of this reference set is critical.

We choose this particular approach for two reasons. First, the leverage of the approach is clearly linked to the strength of the researcher's prior *theoretical* knowledge about the selection

and outcome process. Theoretical knowledge informs the choice of covariates for the reference set, which in turn, conditions the ability (or inability) of the sensitivity analysis to rule out false positive results. Stronger theoretical knowledge makes for stronger sensitivity analysis, and our choice of sensitivity approaches is influenced by a desire to put this relationship front and center.

Second, the approach is easily implementable for applied research. It requires recovered quantities from only a few basic regressions. We include the details of a general Stata package, *poet*, that implements the approach in wide array of settings. Sensitivity analysis is a vibrant field, and our goal is not an exhaustive characterization of all sensitivity tests or advocacy for one "best" approach.[34] Rather, we hope this lowers the barriers to using sensitivity analysis in applied research, while still retaining an emphasis on the precise relationship between theory and the claims being made with the statistical quantity.

We first present the approach, highlighting the issue of choosing covariates for the reference set. We then compare the approach to other well-known approaches from Imbens[35] and Rosenbaum[36], showing their similarities and differences. Lastly, we use two replications from the WTO/CITES replications and one new replication to show how the approach screens likely false positives, upholds likely true positives, and how reference set choices matter.

The Altonji et al Approach

This 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. Note that we use the pairs "ratification/compliance" and "treatment/outcome" interchangeably.

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

$$\text{plim}\hat{\alpha} = \alpha + \frac{\text{var}(r_{it})}{\text{var}(\tilde{r}_{it})} \left[E[u_{it}^c | r_{it} = 1] - E[u_{it}^c | r_{it} = 0] \right],$$

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_{it}^c are the disturbances from a regression of ratification on the observables. $\tilde{r}_{it}$ are the disturbances from a regression of ratification on the observables. $E[u_{it}^c | r_{it} = 1] - E[u_{it}^c | r_{it} = 0]$ describes the degree of selection on unobservables. It is the shift in the distribution of unobservables affecting compliance when comparing ratifiers and non-ratifiers. The term $\frac{var(r_{it})}{var(\tilde{r}_{it})}$ is necessary

---

[34] For two recent advances, see (Blackwell, 2014) on confounding functions and (Imai, Keele, & Yamamoto, 2010) (Imai, Keele, Tingley, & Yamamoto, 2011) on mediation analysis. We also do not cover approaches based on bounds, e.g. (Manski, 1990) (Mebane & Poast, 2013).

[35] Imbens (2003)

[36] Rosenbaum (2002)

18   CHAUDOIN, HAYS, HICKS

to adjust the bias expression after making treatment and the observables orthogonal. Under the null hypothesis of no ratification effect, i.e. $\alpha = 0$, this expression implies Equation 2:

$$E[u_{it}^c | r_{it} = 1] - E[u_{it}^c | r_{it} = 0] = \hat{\alpha} \frac{\text{var}(\tilde{r}_{it})}{\text{var}(r_{it})} \tag{2}$$

The left hand side represents degree of selection on unobservables necessary to explain all of the estimated ratification effect. Is it plausible that the selection problem is this severe? The innovation Altonji et al.[37] is to use "the degree of selection on observables as a guide to the degree of selection on unobservables (Altonji et al., 2005, p. 153)." We start by assuming that selection on unobservables is the same as selection on observables.[38] Formally, this means $\phi_{X'\beta} = \phi_{u_{it}^c}$ in the linear projection of r_{it}^* onto $X'\beta$ and u_{it}^c

$$\text{Proj}(r_{it}^* | X'\beta, u_{it}^c) = \phi_0 + \phi_{X'\beta} X'\beta + \phi_{u_{it}^c} u_{it}^c,$$

where r_{it}^* is the latent variable that determines ratification, $r_{it} = 1(r_{it}^* > 0)$, and β and u_{it}^c are the vector of coefficients and disturbances respectively from a regression of c on X. In other words, the part of the compliance outcome that is attributable to observables, $X'\beta$, has the same marginal effect on selection into the treatment as the part of the compliance outcome that is attributable to unobservables, u_{it}^c. Altonji et al.[40] show that the condition $\phi_{X'\beta} = \phi_{u_{it}^c}$ implies

$$\frac{E[u_{it}^c | r_{it} = 1] - E[u_{it}^c | r_{it} = 0]}{\text{var}(u_{it}^c)} = \frac{E[X'\beta | r_{it} = 1] - E[X'\beta | r_{it} = 0]}{\text{var}(X'\beta)}, \tag{3}$$

which is used to calculate the sensitivity ratio.[41] Substituting Equation 3 into Equation 2 gives us the ratio of selection on unobservables to observables, necessary to drive the effect of ratification to zero:

$$\frac{\hat{\alpha}\text{var}(\tilde{r}_{it})\text{var}(X'\beta)}{\text{var}(r_{it})\text{var}(u_{it}^c)(E[X'\beta | r = 1] - E[X'\beta | r = 0])} \tag{4}$$

In practice, calculating this ratio involves recovering quantities from simple regressions. First, the residuals $\tilde{r}_{it}$ are recovered from regressing ratification on the observables. Second, regressing compliance on those residuals and the observables yields $\hat{\alpha}$. Third, estimating a constrained equation where the effect of ratification is constrained to equal zero (e.g. regressing compliance on the observables, but not ratification) yields u_{it}^c and β.

Covariate Choice

Choosing which observable covariates to include in the reference set is highly consequential, yet this decision has received relatively little attention. Altonji et al.[42] assume that all observable covariates will be related to both the treatment and control and therefore included in the

---

[37] Altonji et al. (2005)

[38] 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[39]. To the extent that covariates are chosen to minimize omitted variable bias in the estimated effect of ratification, this condition will be conservative.

[40] Altonji et al. (2002)

[41] Note that multiplying the numerators of the ratios in this equality by $\text{var}(r_{it})$ makes them covariances with the binary ratification variable. Thus, this condition implies that the marginal effect of observables in a linear probability model of ratification is the same as the marginal effect of unobservables.

[42] Altonji et al. (2005)

reference set. $X'\beta$ thus includes all covariates, so there is no real choice to be made. Related sensitivity approaches, such as Imbens[43], primarily make multiple covariate by covariate comparisons rather than using a single linear combination of observables, so again there is no need to choose. Instead, the emphasis in the extant literature in on the statistics used for benchmarks.

In applied political science research, it will often be the case that making a single comparison with a linear combination of all the covariates is inappropriate, because some of the covariates are not theoretically linked to treatment. Including variables that are not theoretically linked to the treatment produces sensitivity ratios that have little power to detect false positives. This is because the sensitivity tests depend crucially on the relationship between the variance of the linear combination of observables and the conditional expectation of observables across treatment and control groups. Looking at Equation 3, including irrelevant covariates in the reference set, i.e. covariates that are orthogonal to the treatment, does not affect the right hand side numerator since these covariates are balanced across treatment and control groups. However, including them will likely increase the denominator. This artificially inflates the sensitivity ratio, because Equation 3 is inverted when substituted into Equation 2 to yield Equation 4. This raising of the ratio makes the researcher more likely to conclude that the estimated effect is robust, because it has given her less power to screen false positives. The researcher must take care to exclude theoretically irrelevant covariates from the reference set. At a minimum this typically means stripping away a-theoretical trend, lag and dummy variables from the analysis.[44] Sensitivity analysis can be a useful tool for screening false positives and establishing that true positives are robust when implemented carefully, and this requires scrutinizing the set of observable covariates included in the reference set. Otherwise, it can lead researchers astray.

Using a subset of the observable covariates in the reference set also requires a change to the Altonji et al.[45] approach. One needs to condition the quantities in the standardized selection ratio for observables on the covariates excluded from the reference set. More formally, if we divide the covariates into a set that determines both the treatment (ratification) and the outcome (compliance), $\mathbf{X}_j$, and a set that determines the outcome (compliance) only, $\mathbf{X}_k$, the selection on observables ratio from Equation 3 becomes:

$$\frac{E\left(\mathbf{X}_j{}'\hat{\beta}_j|\mathbf{X}'_k, r_{it}=1\right) - E\left(\mathbf{X}_j{}'\hat{\beta}_j|\mathbf{X}'_k, r_{it}=0\right)}{\text{var}\left(\mathbf{X}_j{}'\hat{\beta}_j|\mathbf{X}'_k\right)}.$$

The numerator is obtained from a regression of $\mathbf{X}_j{}'\hat{\beta}_j$ on $\mathbf{X}_k$ and r_{it} while the denominator is taken from a regression of $\mathbf{X}_j{}'\hat{\beta}_j$ on $\mathbf{X}_k$ only. [46]

To be clear, we are not saying that researchers should exclude covariates that are orthogonal to the treatment from the outcome regression. Without these covariates, the estimated treatment

---

[43] Imbens (2003)

[44] If variables are included in the outcome equation to address temporal and (or) spatial dependence, such as deterministic trends and regional dummy variables, both the observable and unobservable covariates, as well as the treatment, have been purged of these relationships with the outcome. Any bias in the estimated treatment effect cannot be attributable to common trending or spatial clustering in unobservables, and therefore it would be misleading to include imbalances across these variables in the sensitivity ratio.

[45] Altonji et al. (2005)

[46] Oster[47] makes a similar point about controls that have no theoretical relationship to unobservable confounds. In her example, gender is an important control variable in a wage regression with education as the treatment, but it has no theoretical relationship to unobserved confounds such as ability and motivation, and therefore should be excluded from sensitivity analysis.

20   CHAUDOIN, HAYS, HICKS

effect would be inefficient. Moreover, leaving the orthogonal covariates out of the outcome regression would bias the estimated marginal effects for the remaining control variables, and this would render the sensitivity ratio uninterpretable. We are suggesting that, after estimating the treatment and covariate effects in a regression that includes the full set of controls, it is important to be careful selecting the covariates (and corresponding marginal effect estimates) that are used to calculate the sensitivity ratio. In other words, there are two steps in this sensitivity analysis. The first step is estimating treatment and marginal effects. The second step is calculating the sensitivity ratio. Covariate selection only applies to the second step.

Relation to Imbens' and Rosenbaum's Approaches

A related approach compares the partialized explanatory power of observables and unobservables (Imbens, 2003). This approach is better known in political science, so our description is more brief.[48] In this approach, the bias from unobservables or, more typically, a single omitted variable is decomposed into the part that is due to the relationship between the omitted variable and the outcome and the part that is due to the omitted variable and the treatment. These relationships are expressed in terms of partial-R^2 statistics which which are chosen for ease of interpretation.[49] For a given bias in an estimated treatment effect, there is a negative relationship between the two partial-R^2 statistics. When one of these sources of bias increases the other must decrease in order to hold the overall bias constant. For purposes of comparison, Imbens[50] generates iso-curves that plot the relationship between the two partial-R^2 statistics, holding the bias constant.

  These two approaches differ in terms of the statistics used to compare observables and unobservables. Altonji et al.[51] use a single marginal effect statistic while Imbens[52] uses two partial-R^2 statistics. But these differences are more apparent than real. We could generate iso-bias curves that plot the relationships between the marginal effects of an omitted variable on both the treatment and the outcome and plot observable covariates in this space—that is, use two marginal effect statistics rather than a single marginal effect statistic that takes the relationship with the outcome (the β in $X'\beta$) as given. Likewise, we could take the two partial-R^2 statistics and express the information they provide as a single measure of explanatory power. We could then equate this measure for observables and unobservables and use it to produce a sensitivity ratio. For instance, partialized between-group explanatory power with respect to the outcome (where the groups are the units that receive the treatment and those that are in the control) combines information about the explanatory power of variables with respect to both the treatment and outcome. The condition that equates the partial between-group explanatory power of unobservables and observables is

$$\frac{[E[u_{it}^c|r_{it}=1]-E[u_{it}^c|r_{it}=0]]^2\text{var}(r_{it})}{\text{var}(u_{it}^c)} = \frac{[E[X'\beta|r_{it}=1]-E[X'\beta|r_{it}=0]]^2\text{var}(r_{it})}{\text{var}(X'\beta)} \tag{5}$$

---

[48] See, for example, (Clarke, 2005) and (Clarke, 2009).

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

[50] Imbens (2003)

[51] Altonji et al. (2005)

[52] Imbens (2003)

If we use this equality to create a sensitivity ratio using explanatory power, as long as the ratios $\frac{\text{var}(\tilde{r}_{it})}{\text{var}(r_{it})}$ and $\frac{\text{var}(X'\beta)}{\text{var}(u_{it}^c)}$ do not differ much from one, the choice of statistic, marginal effect or partial between-group coefficient of determination (Equation 5 vs. Equation 3), will not matter in a qualitative sense. Given the same set of observable covariates, if we conclude that selection on unobservables would have to be stronger (or weaker) than selection on observables using one statistic, we will come to a similar conclusion using the other statistic as well.[53] The choice of covariates is frequently more consequential than the choice of statistic, a point that is undemphasized in the literature on sensitivity testing. We demonstrate the importance of theoretically informed covariate selection below.

Also related, Rosenbaum[54] presents an approach to sensitivity analysis for matched observations that benchmarks against the experimental ideal of random assignment, under which all subjects are equally likely to receive the treatment. With this approach, we ask: how much more likely to receive the treatment would the treated subjects have to be before we would change our conclusion about a causal effect (e.g., fail to reject the null hypothesis)? The answer to this question comes in the form of an odds ratio denoted by Γ. If the critical value is $\Gamma = 2$, for example, the treated would have to be twice as likely to receive the treatment as the untreated. The likelihood that the differences between the treated and control subjects can be explained by hidden bias decreases with Γ. This form of sensitivity analysis is similar in spirit to that in Altonji et al. (2005). The main differences are, first, that Rosenbaum's approach works with matching while Altonji et al.'s sensitivity analysis is regression based. And second, the benchmark for Rosenbaum is random assignment while Altonji et al. use observable covariates to benchmark selection on unobservables.

In the matching context, it makes sense to continue using Rosenbaum bounds. The drawback is that the random assignment benchmark may not always be a good gauge of sensitivity. Is it unreasonable to believe that an unobservable trait makes the treated subjects twice as likely to be treated? For example, in international relations research, is it unreasonable to believe a group of states that sign a human rights treaty shares an unobservable commitment to improve or sustain their good human rights practices, and that this commitment made them twice as likely to sign the treaty as the group of states that chose not to sign? It would be helpful to know whether any observable covariates have an effect of this size on the probability of being treated. For instance, if raising GDP per capita by a relatively small amount doubles the odds that a country will sign the human rights treaty, it seems perfectly reasonable to believe that an unobservable confound could explain away the entire treatment effect. In principle, one could use a propensity score regression for this purpose, in which case, the difference between Altonji et al.'s approach to sensitivity analysis and Rosenbaum's should be small.[55]

---

[53] (Oster, 2014) has proposed a method for sensitivity analysis that incorporates both marginal effects and explanatory power, establishing a more formal connection between (Altonji et al., 2005) and (Imbens, 2003). She also develops a formulation for the sensitivity ratio that does not assume the null hypothesis of no treatment effect is true.

[54] Rosenbaum (2002)

[55] Alternatively, one could calculate odds ratios for observed covariates using the Rosenbaum test and use these ratios to benchmark selection on unobservables. In other words, the differences between the Rosenbaum and Altonji et al. approaches are not fundamental, but rather stem from the way these tests are used in applied empirical research.

22   CHAUDOIN, HAYS, HICKS

Government Revenue False Positive

One example from the main replication exercise, from Gerring, Thacker, and Moreno[56], found that 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), and the result is robust to including country fixed effects and the matching approach. This positive relationship is likely spurious. The WTO explicitly limits tariff barriers, and government revenue data include tariffs as a source of revenue. However, governments that join the WTO tend to be less corrupt and better governed, and thus better able to collect revenue. It is possible that these hypothetical sources of selection on unobservables generated the positive result. This result is useful for demonstrating the ability of sensitivity testing to assess the estimated positive effect and for demonstrating the issues related to covariate selection raised above.

To assess the likelihood of a false positive relationship, we calculate sensitivity ratios for three different linear combinations of observables (reference sets) using: 1) all the covariates from the original regression, 2) all the covariates less the trend, lag and dummy variables, and 3) a theoretically informed subset of covariates. Davis and Wilf[57] allows us to draw on theory to choose the third reference set of covariates. They argue that political variables (such as a country's level of democracy) and economic variables (such as the country's per capita GDP) affect which countries join the GATT/WTO. 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 covariates in the Gerring et al study, we select Centripetalism, Democracy stock, GDP per capita, and Population to include in the third reference set. 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.

The results are in Table 5, which provides the quantities required to calculate the sensitivity ratios. The columns in the upper part of the table give the quantities that vary by linear combination. The lower part of the table gives the quantities in the ratios that do not depend on the reference set. We useˆto denote estimates recovered from particular regressions.[58]

The first thing to note is that the choice of reference set matters greatly for whether the estimated effect is deemed robust. The first sensitivity ratio, based on all the covariates, is 1.41, suggesting the GATT/WTO-tax relationship is robust. However, this is based on a linear combination that includes covariates that are not linked theoretically to membership in GATT/WTO.[59] Benchmarking against irrelevant covariates has given us too little power to detect a false positive, and we worry that this ratio is too large.

As expected, when we begin to prune away the irrelevant covariates, the sensitivity ratios become smaller. The second ratio, which excludes the trend, lag and dummy variables produces a borderline sensitivity ratio of .982. The third calculation, using the theoretically relevant set of covariates, produces a sensitivity ratio of .72, which suggests the GATT/WTO-tax relationship is sensitive. More specifically, the null hypothesis of no ratification effect implies an omitted variable bias or, equivalently, an imbalance in unobservables of 2.336. The imbalance in the linear combination of theoretically relevant variables—Centripetalism, Democracy stock, GDP per capita and, Population and Oil—across the treatment and control groups is 3.258. Thus,

---

[56] Gerring et al. (2005)

[57] Davis and Wilf (2011)

[58] The appendix shows the Stata command and output for the tables in this section.

[59] For example, this set includes a spatial lag in tax revenue. There is no reason why tax revenue in a country's neighbors would affect the likelihood that it joins GATT/WTO.

TABLE 5 *Sensitivity Table for Government Revenue False Positive*

	$\hat{E}\left(\mathbf{X}_j{}'\hat{\beta}_j\|\mathbf{X'}_k, r_{it}=1\right) - \hat{E}\left(\mathbf{X}_j{}'\hat{\beta}_j\|\mathbf{X'}_k, r_{it}=0\right)$	$\widehat{\text{var}}\left(\mathbf{X}_j{}'\hat{\beta}_j\|\mathbf{X'}_k\right)$	$E\left(u_{it}^c\|r_{it}=1\right) - E\left(u_{it}^c\|r_{it}=0\right)^a$	Ratio
	(1)	(2)	(3)	(4)
Observables				
Combination(1)	2.332***	42.46	1.659	1.408
(*All covar.*)				
Combination(2)	0.969***	12.30	2.379	0.982
(*Less trends, lags etc.*)				
Combination(3)	0.784***	7.262	3.258	0.717
(*Only theoretically relevant covar.*)				
Other Quantities				
$\hat{\alpha}$	3.684	$\widehat{\text{var}}(\tilde{r}_{it})/\widehat{\text{var}}(r_{it})$		0.634
$\widehat{\text{var}}(\hat{u}_{it}^c)$	30.20	$E\left(u_{it}^c\|r_{it}=1\right) - E\left(u_{it}^c\|r_{it}=0\right)^b$		2.336

Notes. [a]Imbalance implied by the assumption that selection on observables is equal to selection on unobservables. [b]Imbalance implied by the assumption that the null hypothesis of no treatment effect is true. The sensitivity ratio is the ratio of the latter (b) to the former (a). *** statistically significant at the .01 level.

selection on unobservables (i.e., the imbalance in unobservables) would only have to be 0.72 as strong as selection on the relevant observables (i.e., the imbalance in the relevant observables) to account for the entire estimated treatment effect. This seems plausible. The five variables we identified have a theoretical relationship with WTO membership, but they do not explain all of the variation WTO membership. It is very possible that one or more unobservables are approximately seven-tenths as strong at explaining WTO membership as the observables we used here. In general, the value one marks an important threshold for interpreting sensitivity ratios. A ratio less than one tells us that an imbalance in unobservables across the treatment and control groups that is smaller than the imbalance in the linear combination of theoretically relevant observables would be sufficient to produce an omitted variable bias large enough to account for the entire estimated treatment effect. A ratio greater than one implies that selection on unobservables would have to be stronger than selection on observables in order to explain entirely the estimated treatment effect.

Trade True Positive

How does sensitivity analysis perform in situations where the researcher *does* believe that the institution has an effect? To assess this, we replicated a recent study from Allee and Scalera[60]. The authors argue that some countries who join the WTO face a rigorous, demanding accession process which forces them to make greater concessions and more significant cuts to their pro-

---

[60] Allee and Scalera (2012)

tectionist barriers. Other countries face easier accession processes. They argue that a rigorous accession yields greater subsequent increases to trade. Their dataset uses country-year observations, covering all countries from 1950-2006. They regress the log total trade of country i in year t (the outcome variable) on a dummy variable which indicates whether that country underwent a rigorous GATT/WTO accession (the institutional variable). In addition to period (year) dummy variables, they include five control variables: the log of the country's population, the country's GDP per capita, the number of states bordering the country, democracy, and a measure of internal political conflict. In their main specifications, they find that a rigorous accession yields a 65% increase in total trade, which is statistically significant at the 0.01 level. It is possible that unobservables, such as domestic market structure or factor endowments affect the likelihood of a rigorous accession process and subsequent levels of trade.

Table 6 shows the results of sensitivity analysis. Again we report sensitivity ratios for three linear combinations—(1) all of the variables, (2) all of the variables excluding the period dummies, and (3) a theoretically informed subset of covariates. For the last set, we rely on Pelc[61] who argues that market size and regime type determine the conditions under which countries join the WTO. Thus, we include population, GDP, and Polity.

TABLE 6   *Sensitivity Table for Trade True Positive*

	$\hat{E}\left(\mathbf{X}_j'\hat{\beta}_j\|\mathbf{X}'_k, r_{it}=1\right) -$ $\hat{E}\left(\mathbf{X}_j'\hat{\beta}_j\|\mathbf{X}'_k, r_{it}=0\right)$ (1)	$\widehat{\mathrm{var}}$ $\left(\mathbf{X}_j'\hat{\beta}_j\|\mathbf{X}'_k\right)$ (2)	$E\left(u_{it}^c\|r_{it}=1\right) -$ $E\left(u_{it}^c\|r_{it}=0\right)^a$ (3)	Ratio (4)
Observables				
Combination(1) (*All covar.*)	1.379***	3.800	0.367	1.287
Combination(2) (*Less trends, lags etc.*)	0.639***	2.739	0.236	2.002
Combination(3) (*Only theoretically relevant covar.*)	0.58***	2.346	0.25	1.889
Other Quantities				
$\hat{\alpha}$	0.52	$\widehat{\mathrm{var}}(\tilde{r}_{it})/\widehat{\mathrm{var}}(r_{it})$		0.908
$\widehat{\mathrm{var}}(\hat{u}_{it}^c)$	1.011	$E\left(u_{it}^c\|r_{it}=1\right) - E\left(u_{it}^c\|r_{it}=0\right)^b$		0.472

Notes. [a]Imbalance implied by the assumption that selection on observables is equal to selection on unobservables. [b]Imbalance implied by the assumption that the null hypothesis of no treatment effect is true. The sensitivity ratio is the ratio of the latter (b) to the former (a). *** statistically significant at the .01 level.

We find strong evidence that the rigorous accession treatment effect is robust. The sensitivity ratios for all three linear combinations are greater than one. When we include all of the covariates the sensitivity ratio is 1.29. This implies that the imbalance in unobservables would

[61] Pelc (2011)

have to be almost 30% stronger than the imbalance in observables to account for the entire estimated treatment effect. Our concern with this particular linear combination, as previously, is that it is too large because the reference set of covariates includes variables that are irrelevant for the treatment. Therefore, it is important to narrow the set. However, this time the concern is unwarranted. When we narrow the set to include only theoretically relevant variables, the imbalance in observables decreases at a faster rate than the variance. As a result, the sensitivity ratios get larger rather than smaller.[62] With the theoretically grounded set of covariates we find that selection on unobservables would have to be nearly twice as strong as selection on observables to account for the entire estimated rigorous accession effect.

CONCLUSIONS

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

First, recognizing the problem is inherently important. In the context of IR and international institutions, 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 inference. The context we examined has similarities to many contexts studied in other subfields, where the possibility of false positives also exists. To the best of our knowledge, ours is the first widespread replication exercise to assess the severity of the problem of bias resulting from selection on unobservables.

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. Often, applied researchers lack valid sources of exogenous variation in their explanatory variable, as would be required for an instrumental variables approach or an alternative identification strategy. Even when faced with these problems, sensitivity analysis allows the researcher to assess how sensitive her estimates are to alternative assumptions about the severity of the selection on unobservables problem. Crucially, the leverage generated by the test depends on her theoretical knowledge of the particular context. Theoretical knowledge determines her choice of covariates to include in the implementation of the test, a choice which has large implications for the results and interpretation of the test. Ultimately, the ability of a sensitivity approach to per-

---

[62] This can happen when there are spurious imbalances in covariates with relatively low explanatory power for the outcome. These observables do not provide a useful benchmark for evaluating the plausibility of bias-generating unobservables. The former are nearly irrelevant to both the treatment and the outcome, while bias-generating unobservables are strongly linked to both the treatment and outcome.

26   REFERENCES

suasively screen a false positive and approve a true positive result 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 guidance and tools to leverage their theoretical knowledge in the face of the commonly encountered threat to inference, selection on unobservables.

REFERENCES

Allee, T. L., & Scalera, J. E. (2012). The divergent effects of joining international organizations: Trade gains and the rigors of wto accession. *International Organization*, *66*(02), 243-276.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2002, April). *Selection on observed and un-observed variables: Assessing the effectiveness of catholic schools.* (Manuscript. Dept. Econ., Northwestern Univ.)

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

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. *Handbook of labor economics*, *3*, 1277–1366.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust difference-in-differences estimates? *Quarterly Journal of Economics*, *119*(1), 249-275.

Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, *22*(2), 169-182.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, *22*(4), 341-352.

Clarke, K. A. (2009). Return of the phantom menace: Omitted variable bias in political research. *Conflict Management and Peace Science*, *26*(1), 46-66.

Davis, C., & Wilf, M. (2011). *Joining the club: Accession to the gatt/wto.* (Working Paper, Princeton University)

Downs, G. W., Rocke, D. M., & Barsoom, P. N. (1996). Is the good news about compliance good news about cooperation? *International Organization*, *50*(3), 379–406.

Gerring, J., Thacker, S. C., & Moreno, C. (2005). Centripetal democratic governance: A theory and global inquiry. *The American Political Science Review*, *99*(4), pp. 567-581.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199-236.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*, 765–789.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), pp. 51-71.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, *93*(2), pp. 126-132.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis*, *23*(3), 313-335.

Leuven, E., & Siansei, B. (2003). *Psmatch2: Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.* (Version 4.0.11)

Lupu, Y. (2013). The informative power of treaty commitment: Using the spatial model to address selection effects. *American Journal of Political Science*.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), pp. 319-323.

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

Miller, M. (2015). *The uses and abuses of matching in political science.* (Working Paper)

Oster, E. (2014). Unobservable selection and coefficient stability: Theory and evidence. *University of Chicago Booth School of Business Working Paper*.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press.

Pelc, K. J. (2011). Why do some countries get better wto accession terms than others? *International Organization*, *65*(04), 639–672.

Plumper, T., & Troeger, V. E. (2013). *Not so harmless after all: Fixed effects as identification strategy.* (EPSA Conference Paper)

Rosenbaum, P. R. (2002). *Observational studies.* New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, *12*(1), 487-508.

Simmons, B. A. (2000). International law and state behavior: Commitment and compliance in international monetary affairs. *The American Political Science Review*, *94*(4), 819–835.

Simmons, B. A., & Hopkins, D. J. (2005). The constraining power of international treaties: Theory and methods. *American Political Science Review*, *99*(04), 623-631.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* Springer-Verlag, New York.

Von Stein, J. (2005). Do treaties constrain or screen? selection bias and treaty compliance. *The American Political Science Review*, *99*(4), 611–622.