Do We Really Know the WTO Cures Cancer?
Supporting Information

May 6, 2016

This appendix first describes the Monte Carlo simulations that are mentioned in the main text. This appendix then describes an original Stata command that implements the sensitivity tests described in the main text. We show its syntax and the output from the Gerring et. al. and Allee and Scalera examples in the sensitivity section. Next, we look for patterns in the WTO false positive rates based on the replication study's subject matter, number of countries, and year. We do not find strong patterns. We also show results from Rbounds analysis of the two examples given the most attention in the main manuscript. Then, we list the studies used in the replication exercise. For each study, we give the citation, a brief description of the sample, a description of each variable that was included and the number of replications for each variable. We also note where we excluded some variables from the replications. Third, we list the ratification dates for CITES members and speak briefly about how CITES ratification patterns differ from GATT/WTO ratification patterns.

Simulations and a Generic Data Generating Process

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

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

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

Data-Generating Process with Types of Unobservables

As in the main manuscript, 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 1 the ratification equation and Equation 2 the compliance equation.

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

$$c_{it} = X_{it}\beta + \alpha r_{it} + u^c_{it} \tag{2}$$

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

Unobs. in ratification equation

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

Unobs. in compliance equation

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

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

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

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

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

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

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

Simpler Data-Generating Process and the Law of Second Best

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

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

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

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

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

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

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

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

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

For our continuous compliance experiments, we evaluated the performance of three approaches: OLS without any fixed effects ("do nothing"), unit fixed-effects, and matching, as in Table 1. We use panel-corrected standard errors with our OLS and fixed-effects estimators. For our binary compliance experiments, we evaluated the basic logit and conditional logit estimators and matching approach, as in Table 2. We used time-period clustered and panel-bootstrapped standard errors for our logit and conditional logit estimators respectively. The matching approach is the same as in the replications. Evaluations are based on 1,000 trials.

We expect two trends in the results. First, the false-positive performance of the "do-nothing" estimators should deteriorate across our cases as we move from low to high covariance between the ratification and compliance disturbances. Second, the relative performance of our fixed-effects estimators should improve in our high covariance cases where some of the overall covariance is attributable to unit effects, but deteriorate when this is not the case.

In the case of the unit fixed effects estimator, this arises because of simultaneity bias. Ratification is endogenous if it covaries with the disturbance in the compliance equation. Fixed-effects estimators potentially reduce this covariance, but they also reduce the variance in the ratification

Table 1: 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%

decisions that is leveraged to estimate their effects on compliance. The bias in the estimated treatment effect depends on both of these. The simultaneity bias increases with the strength of the covariance between ratification decisions and the unobservable determinants of compliance, but it decreases as the variance in ratification decisions increases. The first-best solution is to eliminate *all* of the spurious sources of covariance between ratification and compliance. If this can be done, the 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 (2013) finding that unit-fixed effects strategies may be worse than pooled strategies in the presence of unobserved trending. Clarke (2005) and Clarke (2009) 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.[1]

We find both of these patterns in the results. In both Table 1 (continuous compliance) and Table 2 (binary compliance) the performance of the do-nothing OLS and matching approaches gets progressively worse as we move from Case 1 to Case 4.[2] 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

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

[2] Note that the false positive rates are near 1 because we've focused the DGP on only one covariate and the problem of selection on unobservables. Adding other covariates or varying the sample size would affect the nominal false positive rates, but not the trends as we varied ρ.

Table 2: 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%

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

that determine both ratification and compliance, but does reduce significantly the variance in the ratification decisions, it will make the simultaneity bias worse.

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.

1 *poet*: A Stata Command for Sensitivity

The basic syntax for poet[3] is:

poet *depvar* [*theor. relevant indepvars*], treat(*treatmentvar*) other(*other indepvars*)

depvar is the outcome/dependent variable.

theor. relevant indepvars is the reference set of independent variables that are theoretically relevant

[3]Because sensitivity isn't just for poets and musicians.

for selection and outcome. These are the variables whose strength will be used for the benchmark of selection on observables.

treatmentvar is the binary variable indicating treatment (ratification in our examples).

other indepvars is the set of independent variables that are not theoretically related to selection.

Here is an example of the *poet* syntax and output for the Allee and Scalera sensitivity test using the reference set of theoretically relevant variables (combination 3 in the manuscript):

```
.         poet lnFtrade lnpop1 gled_gdppc polity , treat(rigorous) other(totalcont domestic1_9 yr* )

Sensitivity analysis
```

	Obs. Cond. Mean	Cond. Var	Unobs. Cond. Mean	Implied Ratio Altonji
Lin. combination	.58003	2.3456	.25006	1.88873
Alpha hat	.52032			
Var. disturb.	1.0112			
Adj. ratio	.90771			
Selection on obs.	.4723			

All of the quantities shown in the table are also stored as rclass scalars, viewable by typing return list after running the poet command.

2 Patterns in the False Positives

One reasonable question to ask about the replications is: are there patterns in the false positives? In other words, are certain types of studies more or less likely to generate a false positive result for the WTO exercise?

First, we checked if the false positive rates varied by the subject matter. We classified the original studies' dependent variables into four categories: Domestic Economy (e.g. growth rate, infant mortality), IPE (e.g. IMF loans), Political (e.g. change in polity score, labor laws) and Violence (e.g. civil war, torture). Then we calculated the false positive rates for each category. The results are displayed in Table 3. The false positive rates were highest in the economic categories. To double check that our results weren't solely driven by those categories, the final two rows exclude the Domestic Economy and Domestic Economy + IPE categories. False positive rates are still too high, at 28% and 21% respectively.

Second, we checked whether studies that had a global sample of countries, versus a regional focus, were more likely to have false positives. There was actually only one study that focused solely on one region. The rest had a global focus.

Third, we checked whether the number of countries in the study affected the probability of a false positive. We regressed the GATT/WTO p-value on the number of countries (OLS). The coefficient estimate was .0008301 with an associated p-value of 0.207. There does not seem to be a strong relationship between the number of countries and the false positive rates.

Fourth, we checked whether the number of years covered by the study affected the probability of a false positive. We regressed GATT/WTO p-value on the number of countries (OLS). Again, there did not appear to be a strong relationship. The coefficient estimate was .0026296 with an associated p-value of 0.227.

Table 3: False Positive Rates by DV Type

Variable	**Mean**	**Std. Dev.**	**N**
Full Sample	0.33	0.473	94
Domestic Economy	0.667	0.492	12
IPE	0.563	0.512	16
Political	0.211	0.413	38
Violence	0.214	0.418	28
Excl. Dom. Econ.	0.28	0.452	82
Excl. Dom. Econ. & IPE	0.212	0.412	66

In sum, there do not appear to be strong patterns in the false positive rates. This lends support to the idea that the potential for a false positive is widespread in applied work.

An anonymous reviewer asked for a comparison between the Altonji et al approach and more familiar Rbounds analysis for the two examples that we covered in greater detail in the paper. This was a good suggestion. For the Gerring et al false positive (that WTO membership increases tax revenue), Rbound analysis concludes that this result is likely very robust. Even a Γ of 10 is insufficient to raise the upper bound of the significance level for the treatment effect above 0.10.

On the other hand, Rbound analysis concluded that the Allee and Scalera true positive (rigorous WTO accession increases trade) was less robust. A Γ of 3 is sufficient to raise the upper bound to 0.95. Both of these Rbounds results are in Table 4 below. A full comparison of sensitivity approaches is beyond the scope of this paper, but we did find it interesting that the Altonji et al approach gave results that seemed more likely to be accurate, given theoretical knowledge.

Table 4: RBounds results for Gerring et al and Allee and Scalera replications

	Gerring et al	
Γ	sig UB	sig LB
8	0.00055	0
9	0.011728	0
10	0.08379	0

	Allee and Scalera	
Γ	sig UB	sig LB
1	0	0
2	.008512	0
3	.954374	0

3 Studies and Variables Used in Replication Exercise

1. Citation: "Capital, Trade, and the Political Economies of Reform." Sarah M. Brooks and Marcus J. Kurtz. American Journal of Political Science, Vol. 51, No. 4 (Oct., 2007), pp. 703-720.

   - Sample Description: "We examine the dynamics of trade and capital account liberalization in the 19 countries of the Latin American region for which data were available over the period 1985- 99" (710). Approx. 218-221 observations per specification.
   - DV 1: ka open.
     - "The measure of capital account openness is taken from Chinn and Ito (2002), and the reported value is the first standardized principal component for four indicators of capital account regulation: the use of multiple exchange rates, restrictions on cur- rent and capital account, and the compulsory turnover of export receipts. The variable is bounded at roughly -1.8 at the low end and 2.7 at the upper end, with a higher number indicating a more open capital account." (710)
     - 3 specifications/replications.
   - DV 2: trade index. Did not use.

2. Citation: "What Stops Torture?" Courtenay R. Conrad and Will H. Moore. American Journal of Political Science, Vol. 54, No. 2 (Apr., 2010), pp. 459-476.

   - Sample Description: "the spatio-temporal domain covers all countries with populations over one million from 1981 through 1999" (465). Approximately 252 observations.
   - DV 1: no tort.
     - o "In our firstmodel, the 'event' of interest is the termination of torture. Our data to test the aforementioned hypotheses consist of spells of torture in discrete time format. The data cover the years 198199, and there are 3,587 country-years in the data. Of those, 2,341 countryyears are 'at risk' to stopping torture (i.e., that many cases were engaged in torture in the preceding year). The data include 146

countries, and there are a total of 284 spells of torture (i.e., countries that either had a reported use of torture in 1981 when the data set begins or had a reported use of torture in a year after 1981 after having had at least one year in which no uses of torture were uncovered)" (467).

- 3 specifications/replications.

3. Citation: "Master or Servant? Common Agency and the Political Economy of IMF Lending." Mark Copelovitch. International Studies Quarterly. International Studies Quarterly (2010) 54, 4977.

   - Sample Description: "The dataset includes 197 non-concessional IMF loans (Stand-by arrangements and Extended Fund Facility loans, including drawings on the Supplemental Reserve Facility) t‘ 47 countries from 1984 to 2003.1 This sample constitutes the universe of Fund loans during this period by all countries not eligible for the IMFs concessional lending facilities (The Poverty Reduction and Growth Facility and its predecessors). Data on the characteristics of each loan are taken directly from IMF archival documents, including the Letter of Intent declaring a country‘s intent to enter into a Fund program, the corresponding Memorandum of Economic Policies detailing the loan‘s terms and conditionality, and the Staff Reports to the Executive Board." (From the accompanying codebook)..
   - DV 1: amtgdp.
     - Loan size / log(GDP)
     - 1 specifications/replications.
   - DV 2: amtqta.
     - Loan size / log(quota)
     - 2 specifications/replications.
   - DV 3: amtsdr.
     - Loan size / log(millions of SDR)
     - 1 specifications/replications.
   - DV 4: Imfloan.
     - A dummy indicating a new IMF loan for that country-year
     - 1 specifications/replications.
   - DV 5: Bt.
     - Total number of benchmarks/indicative targets
     - 2 specifications/replications.
   - DV 6: pa.
     - Number of performance actions
     - 2 specifications/replications.
   - DV 7: pc.
     - Number of performance criteria
     - 2 specifications/replications.

4. Citation: "Elections and Democratization in Authoritarian Regimes." Daniela Donno. American Journal of Political Science. Vol. 57, No. 3, July 2013, Pp. 703716.

    - Sample Description: I constructed a dataset of elections in EA regimes from 1990 to 2007. The units of analysis are national elections, both presidential and legislative. Creation of the dataset proceeded in several steps. First, Geddes, Wright, and Franz's (2012; GWF hereafter) data were used to identify the set of autocratic regime-years. These data, which cover all countries in the world from 1946 to 2010, expand and improve upon Geddes (1999) earlier coding.7 Within this set, I define the sample of electoral authoritarian regimes as those in which multiple parties are allowed to contest elections. If the ruling party or candidate won less than 75 percent of the votes (presidential contests, first round) or seats (legislative elections) in the last election, the regime is classified as a CAR; otherwise, it is classified as a HAR." (707). Generally, 176-7 observations.
    - DV 1: ed trans.
        - "The dependent variable of the analysis is a transition to electoral democracy. I employ a dummy variable indicating whether a country moved from a "0" to a "1" on Freedom House's list of electoral democracies in the year of the election in question." (708)
        - 5 specifications/replications.

5. Citation: "Centripetal Democratic Governance: A Theory and Global Inquiry." John Gerring, Strom C. Thacker and Carola Moreno. The American Political Science Review, Vol. 99, No. 4 (Nov., 2005), pp. 567-581.

    - Sample Description: Minimally democratic countries from 1960-2000 "Because the theory of centripetalism is applicable only within a democratic framework, we limit all regression analyses to country-years that are minimally democratic, as discussed earlier. Resulting samples vary from a minimum of 77 countries to a maximum of 126, and from a minimum of 14 years to a maximum of 4 decades (1960-2000)" (pg 574)..
    - DV 1: Bureaucratic Quality PRSF.
        - "a measure of political development, is an indicator ranging from 0 to 6 (with higher scores indicating higher quality) developed by the Political Risk Services (PRS) group as part of its International Country Risk Guide (ICRG)" (pg. 573).
        - 2 specifications/replications.
    - DV 2: Life Expectancy ipolate lnF.
        - "measures the expected tenure of life in a country at birth, extrapolating from mortality statistics available at that time (Bos, Vu, and Stephens 1992; Riley 2001; logarithm, data source: World Bank 2003). Like IMR, life expectancy is an overall measure of human development strongly influenced by government policies; hence, it provides a good indicator of the quality of governance in a country." (574).
        - 2 specifications/replications.
    - DV 3: IMR ipolate lnF.
        - "is measured by the infant mortality rate (IMR), the number of deaths per one thousand lives births that occur in the first year of life (logarithm, data source: World Bank 2003). IMR, a primary measure of human development, is affected by many government policies (particularly social policies) and is thus an important outcome-based measure of good governance." (574).

  - 2 specifications/replications.
- DV 4: Illiteracy imputed lnF.
  - "is measured as the percentage of people age 15 and older who cannot, with understanding, both read and write a short, simple statement on their everyday life (logarithm, data source: World Bank 2003). Literacy has become a standard feature of human development indices in recent decades and largely reflects the success of government-sponsored education policies." (574).
  - 2 specifications/replications.
- DV 5: Tax Revenue WDIF.
  - "The variable employed here, drawn from the World Bank's World Development Indicators (World Bank 2003), measures aggregate tax revenues, considered as a share of GDP. More specifically, it counts compulsory, unrequited, nonrepayable receipts for public purposes collected by the central government, including interest collected on tax arrears and penalties collected on nonpayment or late payments of taxes" (pg 573).
  - 2 specifications/replications.
- DV 6: Risk EuromoneyF.
  - "Euromoney ratings are based on polls of economists and political analysts and supplemented by quantitative data such as debt ratios and access to capital markets. The overall country rating derives from nine separate categories, each with an assigned weighting (in parentheses): (1) political risk (25%); (2) economic performance (25%); (3) debt indicators (10%); (4) debt in default or rescheduled (10%); (5) credit ratings (10%); (6) access to bank finance (5%); (7) access to short-term finance (5%); (8) access to capital markets (5%); and (9) discount on forfeiting (5%) (Euromoney 2004)" (pg 573).
  - 2 specifications/replications.
- Didn't use trade openness or GDP per capita.

6. Citation: "Testing the Effects of Independent Judiciaries on the Likelihood of Democratic Backsliding." Douglas M. Gibler and Kirk A. Randazzo. American Journal of Political Science, Vol. 55, No. 3 (July 2011), pp. 696-709.
   - Sample Description: 1960-2000. Generally, about 2,700+ observations.163 countries..
   - "We use two dependent variables to test the effects of in dependent judiciaries within the state, both of which are drawn from the Polity IV (Marshall and Jaggers 2002) dataset. First, we examine any negative changes in overall regime score (autocracy/democracy). Normally, the combined score ranges from 10 to +10. However, since we are using the Henisz data, which rely on the executive constraints composite of Polity IV data, our scale omits executive constraints, and thus ranges from 7 to +6. Consequently, states that are considered democracies on this adjusted scale range from 3 to 6 (instead of 6 to 10 on the Polity scale). The second dependent variable is based on the first. Since the literature on regime reversion depends heavily on crisis mechanisms and regime collapses, we also include analyses of large-magnitude regime changes, which we define as negative changes of 4 points or more.4 Both of these dependent variables are dichotomous and measure only the presence of any negative change or a large-magnitude regime change." 701-702

- DV 1: negpolchg.
  - This is the regime change variable.
  - 4 specifications/replications.
- DV 2: neg4polchg.
  - o This is the regime collapse variable.
  - 4 specifications/replications.

7. Citation: "A Global Model for Forecasting Political Instability." Goldstone et. al. American Journal of Political Science, Vol. 54, No. 1, January 2010, pp. 190208.
   - Sample Description: "We compiled our data from open sources to construct a cross-national time-series data set covering the period 1955 through 2003 for all countries with a population over 500,000. We identified 'instability episodes' in part by identifying conflicts from existing databases (such as the Correlates ofWar) and in part by consulting with area experts." (191). Table 1 N ranges from 196-468.
   - DV 1: sftpcons.
     - Onset of an instability episode
     - 9 specifications/replications.

8. Citation: "Trade-based Diffusion of Labor Rights: A Panel Study, 19862002." BRIAN GREENHILL, LAYNA MOSLEY and ASEEM PRAKASH. The American Political Science Review, Vol. 103, No. 4 (November 2009), pp. 669-690.
   - Sample Description: "We model the relationship between each country's col lective labor rights outcomes and those of its trading partners using country-year data for 90 developing countries over the period 1986-2002. Countries from Africa, Latin America, the Caribbean, Asia, and the Middle East are included in our sample; we exclude the transition economies of Central and Eastern Europe, as well as those from the former Soviet Union. Omitted country-years from the developing regions are those for which data on one or more independent variables are not available." (675).
   - DV 1: lawpos.
     - The first measure, Labor Laws, gives an indication of the extent to which laws have been put in place to safeguard collective labor rights, such as the rights to organize, bargain collec tively, and strike. (675)
     - 3 specifications/replications.
   - DV 2: practicepos.
     - The second measure, Labor Practices, provides an indication of the degree to which labor rights are violated in practice. (675)
     - 3 specifications/replications.

9. Citation: "Do Economic Sanctions Destabilize Country Leaders?" Nikolay Marinov. American Journal of Political Science, Vol. 49, No. 3 (Jul., 2005), pp. 564-576.

- Sample Description: "I include in the test all countries with population over 500,000 (N = 160), and the period 1947 to 1999. This makes for a total of 6,782 observations in the full data-set." (568).
- DV 1: fail.
  - Leadership change. "The principal source of data on government leaders is Chiozza and Goemans (2004b), as updated by Goemans et al. (2004)." (569)
  - 3 specifications/replications.

10. Citation: "Foreign Aid Shocks as a Cause of Violent Armed Conflict." Richard A. Nielsen, Michael G. Findley, Zachary S. Davis, Tara Candland and Daniel L. Nielson. American Journal of Political Science, Vol. 55, No. 2 (April 2011), pp. 219-232.

    - Sample Description: 139 countries, 1981-2005. N ranges from 698 to 2627 in main Table..
    - DV 1: prio.
      - UCDP PRIO civil war data, binary indicator for 25 or more battle deaths from civil war
      - 2 specifications/replications.

11. Citation: "Foreign Direct Investment, Regime Type, and Labor Protest in Developing Countries." Graeme B. Robertson and Emmanuel Teitelbaum. American Journal of Political Science, Vol. 55, No. 3 (July 2011), pp. 665-677.

    - Sample Description: 129-131 countries, I think 1980-2005.
    - DV 1: dispute.
      - "For our analysis, we use data from the High Pro file Strikes Dataset (HPSD), which allows for the analyst to distinguish among a greater number of strike charac teristics and corrects for some of the reporting bias of available measures. The data for the HPSD come from press reports, which were gathered using broad search terms (e.g., 'labor' and 'strikes') to gather all articles in the 'World Publications' section of the Nexis database documenting political or economic trade union protest in all countries that could be categorized as 'non-OECD countries' as of 1980. This process yielded 1,069 protest events in 84 countries from 1980 to 2005, of which 603 can be clearly defined as industrial disputes and 351 as political strikes." (670)
      - 5 specifications/replications.

12. Citation: "Refugees and the Spread of Civil War." Idean Salehyan and Kristian Skrede Gleditsch. International Organization. Volume 60, Issue 02, April 2006, pp 335 366.

    - Sample Description: 1950-2001, country years, but dropping subsequent onsets (350). N = 5567 in the main table.
    - DV 1: nonset.

  - "Our conflict data come from the Uppsala0PRIO Conflict Data Set These data identify instances of armed conflict involving more than twenty-five casualties in a given calendar year+ As a robustness check, we also reestimate our model, restricting the analysis to more severe wars involving at least 1,000 battle deaths over the course of the conflict For our dependent variable, we include data on intrastate and internationalized intrastate conflicts where a state experiences conflict on its own territory, as classified by the location variable in the Uppsala PRIO data set Our main dependent variable is conflict onset, which is coded 1 for the first year of a conflict and 0 if no conflict takes place in the state in that particular year+ Subsequent ongoing years of the same conflict are dropped from the estimation sample In cases where there were multiple conflict onsets in a country, data on a new onset was included if it occurred during the years when another conflict was ongoing." (350)
  - 5 specifications/replications.
- DV 2: bigconset.
  - 1 specifications/replications.

13. Citation: "Foreign Aid, Democratization, and Civil Conflict: How Does Democracy Aid Affect Civil Conflict?" Burcu Savun and Daniel C. Tirone. American Journal of Political Science, Vol. 55, No. 2 (April 2011), pp. 233-246.

    - Sample Description: "The sample for our study is composed of Official Devel opment Aid (ODA) eligible countries between 1990 and 2003.6 There has been a steady increase in the number of democracy aid recipient countries over the years. While only 30 countries received OECD democracy aid in 1990, this number increased to 76 in 1995, and 134 countries received democracy aid in 2003. The unit of analysis is country-year." (237) N=1478-1600 in the main tables.
    - DV 1: Fconflict initiation.
      - "The dependent variable is Conflict Initiation, a dummy variable assuming a value of 1 for a given year if a domestic conflict with at least 25 battle deaths be gins after at least two years without an initiation. We use the UCDP/PRIO Armed Conflict dataset for this variable (Gleditsch et al. 2002)."
      - 5 specifications/replications.

14. Citation: "Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention against Torture." James Raymond Vreeland. International Organization, Vol. 62, No. 1 (Winter, 2008), pp. 65-101.

    - Sample Description: "What is observed regarding patterns of torture under dictatorship? The Hathaway torture data include 967 country-year observations of from 1985 to 1996 covering 109 separate dictatorships. The mean rate of torture is 3.0 with a standard deviation of 1.1; the median is 3." (80) N = 428-694 in the main table..
    - DV 1: Common tort hath.
      - "What is observed regarding patterns of torture under dictatorship? The Hathaway torture data include 967 country-year observations of from 1985 to 1996 covering 109 separate dictatorships.53 The mean rate of torture is 3.0 with a standard deviation of 1.1; the median is 3." (80).

  - 3 specifications/replications.

15. Citation: "Do Authoritarian Institutions Constrain? How Legislatures Affect Economic Growth and Investment." Joseph Wright. American Journal of Political Science, Vol. 52, No. 2, April 2008, Pp. 322343.

    - Sample Description: "To test the preceding hypotheses, an updated version of Geddes (1999) data on authoritarian regime types is used. The updated data include monarchies and author See the web appendix for updated regime type (A) and legislatures coding (B), with a brief review of the coding rules. itarian regime-years for regimes that lasted less than four years. I then updated Przeworski and colleague's (2000) data on authoritarian legislature and parties through 2002. As Table 4 shows, legislatures are present in 69% of personalist regime-years, 62% of monarchy regimeyears, 92% of single-party regime-years, but only 37% of military regime-years. While all regimes are more likely to have legislatures in the postCold War period, significant variation exists in the dependent variable during this period." (329).
    - DV 1: growth.
      - Table 7
      - 2 specifications/replications.
    - DV 2: InvestGDP.
      - DV in table 6.
      - 2 specifications/replications.
    - DV 3: legislature.
      - DV in table 5.
      - 1 specifications/replications.
    - There were tons of specifications. We didn't use all of them, to make sure this paper wasn't massively over-represented.

16. Citation: "How Foreign Aid Can Foster Democratization in Authoritarian Regimes." Joseph Wright. American Journal of Political Science, Vol. 53, No. 3, July 2009, Pp. 552571.

    - Sample Description: "To test the preceding hypotheses, I use an updated version of Geddes's (2003) data on authoritarian regimes. (Wright 2008). The original data were updated by including monarchies, post-Soviet regimes in Central Asia, and authoritarian regime-years for regimes that lasted fewer than three years. The updated data are grouped into four main types of authoritarian regimes: military, monarchy, personalist, and single party." (556-557).
    - DV 1: dem.
      - Tranisition to democracy, Table 2.
      - 4 specifications/replications.
    - DV 2: dpol3.
      - Change in polity, table 3.
      - 4 specifications/replications.

4 CITES Ratification Dates

The most important feature of the CITES replications is that there is unlikely to be a strong theoretical link between CITES and the replication dependent variables. Another good feature is that CITES ratification patterns are different from GATT/WTO ratification/accession patterns. Large, liberal democracies tend to be the fastest GATT/WTO joiners, which is a potential explanation for some of the false positive results. This is not the case with CITES. There is a diverse group of joiners at each stage, mainly because the treaty is based on endangered species protection. Below, we list the ratification and accession dates for the CITES countries. This list contains the countries which were included in at least one of the 13 of 16 replication studies that used Correlates of War country codes. This information can be found at `http://www.cites.org/eng/disc/parties/chronolo.php?order=field_country_date_of_joining&sort=as` (Accessed 2-27-2015).

References

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

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

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

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

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

Table 5: CITES Ratification Dates

country	iso	type	join_day	join_month	join_year
United States of America	US	Ratification	14	1	1974
Nigeria	NG	Ratification	9	5	1974
Switzerland	CH	Ratification	9	7	1974
Tunisia	TN	Ratification	10	7	1974
Sweden	SE	Ratification	20	8	1974
Cyprus	CY	Ratification	18	10	1974
Ecuador	EC	Ratification	11	2	1975
Chile	CL	Ratification	14	2	1975
Uruguay	UY	Ratification	2	4	1975
Canada	CA	Ratification	10	4	1975
Mauritius	MU	Ratification	28	4	1975
Nepal	NP	Accession	18	6	1975
Peru	PE	Ratification	27	6	1975
Costa Rica	CR	Ratification	30	6	1975
South Africa	ZA	Ratification	15	7	1975
Brazil	BR	Ratification	6	8	1975
Madagascar	MG	Ratification	20	8	1975
Niger	NE	Ratification	8	9	1975
Morocco	MA	Ratification	16	10	1975
Ghana	GH	Ratification	14	11	1975
Papua New Guinea	PG	Accession	12	12	1975
Germany	DE	Ratification	22	3	1976
Pakistan	PK	Accession	20	4	1976
Finland	FI	Accession	10	5	1976
DR Congo	CD	Accession	20	7	1976
India	IN	Ratification	20	7	1976
Norway	NO	Ratification	27	7	1976
Australia	AU	Ratification	29	7	1976
United Kingdom	GB	Ratification	2	8	1976
Iran	IR	Ratification	3	8	1976
Paraguay	PY	Ratification	15	11	1976
Guyana	GY	Accession	27	5	1977
Denmark	DK	Ratification	26	7	1977
Senegal	SN	Accession	5	8	1977
Nicaragua	NI	Accession	6	8	1977
Gambia	GM	Accession	26	8	1977
Malaysia	MY	Accession	20	10	1977
Venezuela	VE	Ratification	24	10	1977
Botswana	BW	Accession	14	11	1977
Egypt	EG	Accession	4	1	1978

Table 6: CITES Ratification Dates, Continued

France	FR	Approval	11	5	1978
Panama	PA	Ratification	17	8	1978
Togo	TG	Ratification	23	10	1978
Kenya	KE	Ratification	13	12	1978
Jordan	JO	Accession	14	12	1978
Indonesia	ID	Accession	28	12	1978
Sri Lanka	LK	Accession	4	5	1979
Bahamas	BS	Accession	20	6	1979
Bolivia	BO	Ratification	6	7	1979
Italy	IT	Ratification	2	10	1979
Guatemala	GT	Ratification	7	11	1979
Tanzania	TZ	Ratification	29	11	1979
Israel	IL	Ratification	18	12	1979
Japan	JP	Acceptance	6	8	1980
CAR	CF	Accession	27	8	1980
Rwanda	RW	Accession	20	10	1980
Suriname	SR	Accession	17	11	1980
Zambia	ZM	Accession	24	11	1980
Portugal	PT	Ratification	11	12	1980
Argentina	AR	Ratification	8	1	1981
China	CN	Accession	8	1	1981
Liberia	LR	Accession	11	3	1981
Mozambique	MZ	Accession	25	3	1981
Zimbabwe	ZW	Accession	19	5	1981
Cameroon	CM	Accession	5	6	1981
Philippines	PH	Ratification	18	8	1981
Colombia	CO	Ratification	31	8	1981
Guinea	GN	Accession	21	9	1981
Bangladesh	BD	Ratification	20	11	1981
Austria	AT	Accession	27	1	1982
Malawi	MW	Accession	5	2	1982
Sudan	SD	Ratification	26	10	1982
Thailand	TH	Ratification	21	1	1983
Congo	CG	Accession	31	1	1983
Belgium	BE	Ratification	3	10	1983
Algeria	DZ	Accession	23	11	1983
Luxembourg	LU	Ratification	13	12	1983
Trinidad and Tobago	TT	Accession	19	1	1984
Benin	BJ	Accession	28	2	1984
Netherlands	NL	Ratification	19	4	1984
Honduras	HN	Accession	15	3	1985
Hungary	HU	Accession	29	5	1985

Table 7: CITES Ratification Dates, Continued

Afghanistan	AF	Accession	30	10	1985
Somalia	SO	Accession	2	12	1985
Spain	ES	Accession	30	5	1986
Belize	BZ	Succession	19	8	1986
Singapore	SG	Accession	30	11	1986
Dominican Republic	DO	Accession	17	12	1986
El Salvador	SV	Accession	30	4	1987
Burundi	BI	Accession	8	8	1988
Chad	TD	Accession	2	2	1989
Gabon	GA	Accession	13	2	1989
Ethiopia	ET	Accession	5	4	1989
Malta	MT	Accession	17	4	1989
New Zealand	NZ	Accession	10	5	1989
Burkina Faso	BF	Accession	13	10	1989
Poland	PL	Ratification	12	12	1989
United Arab Emirates	AE	Accession	8	2	1990
Cuba	CU	Accession	20	4	1990
Guinea-Bissau	GW	Accession	16	5	1990
Namibia	NA	Accession	18	12	1990
Bulgaria	BG	Accession	16	1	1991
Mexico	MX	Accession	2	7	1991
Uganda	UG	Accession	18	7	1991
Russian Federation	RU	Continuation	13	1	1992
Djibouti	DJ	Accession	7	2	1992
Estonia	EE	Accession	22	7	1992
Greece	GR	Accession	8	10	1992
Barbados	BB	Accession	9	12	1992
Slovakia	SK	Succession	2	3	1993
Czech Republic	CZ	Succession	14	4	1993
Republic of Korea	KR	Accession	9	7	1993
Viet Nam	VN	Accession	20	1	1994
Mali	ML	Accession	18	7	1994
Romania	RO	Accession	18	8	1994
Eritrea	ER	Accession	24	10	1994
Sierra Leone	SL	Accession	28	10	1994
Cte d'Ivoire	CI	Accession	21	11	1994
Comoros	KM	Accession	23	11	1994
Belarus	BY	Accession	10	8	1995

Table 8: CITES Ratification Dates, Continued

Mongolia	MN	Accession	5	1	1996
Saudi Arabia	SA	Accession	12	3	1996
Georgia	GE	Accession	13	9	1996
Turkey	TR	Accession	23	9	1996
Latvia	LV	Accession	11	2	1997
Swaziland	SZ	Accession	26	2	1997
Jamaica	JM	Accession	23	4	1997
Yemen	YE	Accession	5	5	1997
Myanmar	MM	Accession	13	6	1997
Cambodia	KH	Ratification	4	7	1997
Uzbekistan	UZ	Accession	10	7	1997
Fiji	FJ	Accession	30	9	1997
Mauritania	MR	Accession	13	3	1998
Azerbaijan	AZ	Accession	23	11	1998
Ukraine	UA	Accession	30	12	1999
Iceland	IS	Accession	3	1	2000
Kazakhstan	KZ	Accession	20	1	2000
Slovenia	SI	Accession	24	1	2000
Croatia	HR	Accession	14	3	2000
Fmr. Macedonia	MK	Accession	4	7	2000
Republic of Moldova	MD	Accession	29	3	2001
Lithuania	LT	Accession	10	12	2001
Ireland	IE	Ratification	8	1	2002
Kuwait	KW	Ratification	12	8	2002
Bhutan	BT	Accession	15	8	2002
Libya	LY	Accession	28	1	2003
Syrian Arab Republic	SY	Accession	30	4	2003
Albania	AL	Accession	27	6	2003
Lesotho	LS	Ratification	1	10	2003
Lao	LA	Accession	1	3	2004
Cape Verde	CV	Accession	10	8	2005

Table 9: CITES Ratification Dates, Continued

Solomon Islands	SB	Accession	26	3	2007
Kyrgyzstan	KG	Accession	4	6	2007
Oman	OM	Accession	19	3	2008
Armenia	AM	Accession	23	10	2008
Bosnia and Herzegovina	BA	Accession	21	1	2009
Bahrain	BH	Accession	19	8	2012
Lebanon	LB	Accession	25	2	2013
Angola	AO	Accession	2	10	2013
Iraq	IQ	Accession	5	2	2014