The Constraining Power of International Treaties: Theory and Methods

BETH A. SIMMONS *Harvard University*
DANIEL J. HOPKINS *Harvard University*

We acknowledge the contribution of von Stein (2005) in calling attention to the very real problem of selection bias in estimating treaty effects. Nonetheless, we dispute both von Stein's theoretical and empirical conclusions. Theoretically, we contend that treaties can both screen and constrain simultaneously, meaning that findings of screening do nothing to undermine the claim that treaties constrain state behavior as well. Empirically, we question von Stein's estimator on several grounds, including its strong distributional assumptions and its statistical inconsistency. We then illustrate that selection bias does not account for much of the difference between Simmons's (2000) and von Stein's (2005) estimated treaty effects, and instead reframe the problem as one of model dependency. Using a preprocessing matching step to reduce that dependency, we then illustrate treaty effects that are both substantively and statistically significant—and that are quite close in magnitude to those reported by Simmons.

Serious researchers in political science are finally beginning to take international legal agreements as worthy of sustained and rigorous analysis. Within the last several years, a growing group of scholars is making progress toward understanding the extent to which international law—and most specifically, the highly public and legal form of commitment represented in treaties—can actually shape the decisions governments make as well as broader outcomes of normative concern. The theory these studies draw on is becoming more refined: increasingly scholars are willing to analyze international legal agreements as a specific kind of commitment device. Treaties are the most formal "language" governments have to focus the expectations of individuals, firms, and other states that they seriously intend to keep their word in a particular policy area. Treaties enhance the reputational effects that may inhere in general policy declarations, precisely because they link performance to a broader principle that underlies the entire edifice of international law: *pacta sunt servanda*—treaties are to be observed. By choosing to become a treaty party, governments ante up a greater reputational stake than would otherwise be the case.

Estimating treaty effects is no simple thing, however. Despite terrific progress in supplementing case studies with quantitative models that test the generality of the claim that legal commitments matter, the evidentiary hurdles and methodological issues are highly contested. The most common worry is that treaty effects are merely reflections of underlying state preferences rather than evidence of an independent influence on behavior (Downs, Rocke, and Barsoom 1996). This assertion is indeed troubling for those who would like to believe that governments can be nudged—if only at the margins—toward internationally preferred behaviors by making explicit agreements.

Jana von Stein[1] has made an important and sophisticated statistical contribution in this regard. Her strategy is to adapt a Heckman selection model to reestimate the impact of signing onto Article VIII, the section of the International Monetary Fund's (IMF) Articles of Agreement that prohibits signatories from restricting their current accounts. As a result, she argues that we should revise our estimates of the treaty's effect downward. Although Simmons (2000, 831) found that the marginal effect of signing onto Article VIII can be up to 27 percentage points in the second year after the last restriction, von Stein revises that estimate to 13 percentage points, although the estimated treaty effect remains both substantively and statistically significant.[2] Perhaps unintentionally, she also makes a contribution by showing that Simmons' original findings were sensitive to the strategy used to control for time, a point we develop here.

It would seem natural to apply a Heckman selection model to the potential problem of treaty selection bias. Our rejoinder is primarily cautionary. Choosing to attack selection bias[3] statistically rather than theoretically and empirically may account for selection "problems" without shedding much light on them. In statistical terms, von Stein argues that Article VIII is

Beth A. Simmons is Professor, Department of Government, Harvard University, 1033 Massachusetts Avenue, Cambridge, MA 02139 (bsimmons@latte.harvard.edu).

Daniel J. Hopkins is a Ph.D. student, Department of Government, Harvard University, Littauer Center, North Yard, Cambridge, MA 02138 (dhopkins@fas.harvard.edu).

The authors thank Adam Berinsky, Lynn Gale, Gary King, and Jasjeet Sekhon for statistical advice, three anonymous reviewers for valuable feedback, and Jana von Stein for helpful comments and assistance in replicating her model. All errors are our own. The data and code needed to replicate the analyses are available from the authors via email.

[1] All references to Jana von Stein refer to von Stein's article in this issue of *APSR*.

[2] As compared with Simmons (2000), von Stein also argues that the estimated treaty effects fade more quickly as the time from the most recent restriction passes. However, throughout this response, we emphasize the results in the first few years after the last restriction, as the data show these initial years to be the most critical in setting states on a restriction-free course.

[3] To be precise, by "selection bias," we mean the bias resulting from the nonrandom assignment of the treatment stemming from both observable and unobservable sources.

not randomly assigned to countries even conditional on observed covariates, and that is indeed problematic. Theoretically, of course, this is to state the obvious. Random assignment would imply a theory of frivolous commitment-making, hardly a model on which a useful theory of compliance with legal obligations can be developed. We *know* treaty commitment is not random; that was shown in the original article (Simmons 2000). It does not follow that treaties are ineffectual. We view the process of making a treaty commitment as a costly policy that only a government with intentions to comply would generally be willing to make. *Ex ante*, for most governments, treaties involve ratification costs.[4] Government must have—or assemble—the basic political support to announce a change in legal regime for a particular policy. We should in most cases expect treaty ratification to be more costly *ex ante* than a mere policy announcement, because the ratification coalition will have to include not only those who may support the policy, but also those who want to tie the government's hands through altering the legal (and normative) setting in which policy is carried out. Because treaties focus expectations on compliance, ratifying a treaty without an intent to comply only raises *ex post* consistency costs. Indeed, the anticipation of such *ex post* costs should in fact contribute to the political opposition (hence, costs) a government faces *ex ante*. The bottom line is this: if treaties are commitment devices, then they *should* in fact have a screening effect, because only those governments that are willing and think they will be able to comply should sign on.

It is essential, however, to correct two implications of von Stein's discussion of treaty effects. First is the implication that anticipatory compliance casts doubt on the commitment value of the treaty itself. There is no reason to think this observed behavior undercuts a theory of the constraining power of treaties. Governments should rationally be concerned about the reputational costs of inconsistency. To move toward compliance prior to a formal commitment may reduce the uncertainty surrounding the ability to comply and is perfectly consistent with the theory advanced here. In fact, one shortcoming of the original article might have been to cast treaty effects too narrowly. If we include the anticipatory compliance treaties induce, we are likely to conclude Article VIII has an even more significant impact than Simmons (2000) originally reported.[5]

Second, and even more worrisome, von Stein's discussion suggests that screening effects and constraining effects are somehow mutually exclusive. We disagree. Even for the committed, there may be conditions under which it would be tempting to renege on a treaty commitment. Many of these conditions will not have been fully anticipated by the government or indeed the ratifying coalition. But having paid the *ex ante* costs of ratification, a legally committed government will still rationally want to avoid the inconsistency costs of reneging. Our argument is that, *facing similar conditions*, Article VIII countries will try harder than will uncommitted countries to avoid restrictions, because they have staked their reputations on doing so. Screening and constraining are compatible treaty functions. The only real question is: how can we distinguish these effects empirically? As we show in the next section, the estimator offered by von Stein offers some advantages, but some serious drawbacks as well.

HECKMAN SELECTION MODELS: A SOLUTION TO THE PROBLEM OF SELECTION BIAS?

Jana von Stein offers a potential solution to the problem of selection bias. She assumes that some of the important factors that explain selection into a treaty regime are unobservable and adapts a Heckman selection model to cope with that bias stemming from selection on unobservables. This section takes a close look at that choice. As is well known, Heckman models have some important limitations, and we demonstrate that, in this application, those limitations are pronounced. But we also believe von Stein has not conclusively isolated *selection effects*, and that much of the difference between her estimates and those in Simmons's (2000) original article are due to other specification choices. Having reframed the problem as one of model dependence, we go on to estimate the impact of Article VIII using techniques that markedly reduce model dependence—hence, that render more reliable results.

Generic Issues

Heckman selection models[6] have enjoyed a recent burst of popularity in the political science literature (Berinsky 1999; Lemke and Reed 2001; Reed 2000; Timpone 1998; Vreeland 2003), although political methodologists are well aware of the problems with this class of models (Sartori 2003; Signorino 2003). Research has shown that Heckman-style models share several important weaknesses, including their sensitivity to specification, possible problems of collinearity, and heavy reliance on distributional assumptions (Lee 2001; Liao 1995; Sartori; Winship and Mare 1992). For precisely these reasons, recent methodological work on selection bias has focused on finding alternatives to the Heckman approach, often through semiparametric or nonparametric models (Heckman et al. 1998; Lee 2001; Sartori 2003; Winship and Morgan 1999). We

[4] We are here using "ratification" in its broad political rather than narrow legal sense, although for some countries and issue areas they will be essentially the same.

[5] At the same time, we should also point out that for the subset of observations we employ in our following reestimation, these anticipated effects are far less pronounced than von Stein argues. The results presented in Table 2 are quite representative. When looking at the subset of signatories for which matched nonsignatories are available for the first of the matched datasets, the change in restriction behavior in the 4 years prior to signing is from restricting 70.5% of the time to restricting 65.9% of the time. The other matched datasets produce similar results.

[6] For some of Heckman's initial work on selection bias, see Heckman 1976, 1979.

explore one such alternative, matching, in the final section of this article.

The problem of being overly reliant on distributional assumptions is a real issue here. In cases where the independent variables for the selection and outcome equations are the same, the standard Heckman selection model is identified *solely* on its distributional assumptions (Sartori 2003). To be sure, von Stein's model includes several variables that appear only in the selection equation, but no theoretical justification is given for why any of those variables is related to restriction behavior *only* through its impact on Article VIII commitment. In other words, it is unclear why any of those variables is a valid instrument with which to identify the model. We thus agree with Winship and Mare (1992, 342) who conclude that "Heckman's method is no panacea for selection problems and, when its assumptions are not met, may yield misleading results," a point that is also made by (Lee 1984). The problem of sensitivity to strong assumptions is especially pronounced in the case of von Stein's estimator, as she adds a second assumption of bivariate normality to a model that has already been criticized precisely for its dependence on distributional assumptions.

An Inconsistent Estimator

An additional concern is specific to von Stein's adaptation of the Heckman probit. As she notes, she includes an indicator variable in the selection equation for observations that occur after a country has signed onto Article VIII. According to von Stein (??), the role of this indicator variable is to approximate survival analysis within a probit model by ensuring that the "estimated coefficients . . . are based only on the values of the independent variables before or in year *t*′," where "*t*′" refers to the year of signing. However, this indicator variable violates the non-quasi-complete separation assumption of logit and probit models: for probit models to render consistent estimates, they cannot include any independent variables that are perfect or quasi-perfect predictors of the dependent variable (Albert and Anderson 1984; Christmann and Rousseeuw 2001). Because observations only are coded as a "1" for this indicator variable if they have signed onto Article VIII—and are never coded as "1" when countries have not signed Article VIII—the indicator variable is a quasi-perfect predictor of the dependent variable. In such cases, there is no overlap between those observations that are predicted to be failures and those predicted to be successes; thus, the maximum likelihood estimates for the model's parameters do not exist. Some computer programs report parameter estimates under these conditions, but those estimates are not correct (Christmann and Rousseeuw 2001).

Put differently, the inclusion of this indicator variable means that even asymptotically, von Stein's estimator does not converge to the right estimates. One way to recognize this problem is to see if there are fitted values that only differ from 0 or 1 by tiny margins, and indeed, some 1,232 observations are predicted to sign with a probability above .999999. We confirmed these suspicions using the Noverlap package in R 2.0.1 (R Development Core Team, 2004; Rousseeuw and Christmann 2004), which shows that there is no overlap when the indicator variable for signatories is included. The resulting inconsistency alone should constitute grounds to reject the estimated treaty effects von Stein presents.

Explaining the Difference in Results: Selection Bias or Other Model Dependencies?

We have discussed the generic problems associated with Heckman selection models, and have argued that von Stein's adaptation produces estimates that are statistically inconsistent. Setting these issues aside, we now turn to whether von Stein has made a case that accounting for selection bias leads to a drastic revision of the estimated impact of Article VIII. Simmons (2000) estimated that Article VIII status makes the country on average 27 percentage points less likely to place restrictions, whereas von Stein estimates the treaty's actual *constraining* effect to be just 13 percentage points. von Stein (XX) attributes this gap *entirely* to selection bias: as she explains, "selection bias accounts for between 31% and 95% of the standard probit model's estimated effect of the legal commitment on a state's propensity to engage in compliant behavior." But von Stein has actually made several simultaneous changes to the original model, and only by disentangling them can we truly understand the extent to which Simmons's original estimate was driven by selection bias.

First, von Stein has changed the definition of the causal effect to be estimated. In a standard one-stage model, researchers often estimate causal effects by varying one or more independent variables while fixing the others to some value and then observing the difference in simulated values of the dependent variable under the model (King, Tomz, and Wittenberg 2000). In the model proposed by von Stein, we have separate estimates for the second-stage coefficients of the signatories and the nonsignatories. Instead of varying the values of key independent variables, then, von Stein fixes the values of all explanatory variables and varies the set of coefficients used to calculate the predicted probabilities. But if this is our strategy for estimating predicted probabilities, we need to specify *a priori* which values of the independent variables are of interest. Do we care about the effect of the treatment on the treated population, on the nontreated population, or on some other group?

von Stein chooses to focus on the effect of the treatment on the *nontreated*: she generates her estimates of the impact of Article VIII by measuring the change in the predicted probability that the mean nonsignatory will restrict its current account using first the nonsignatory and then the signatory outcome equations. But she could as easily have chosen to estimate the effect on the *treated* population instead. And in fact, when we estimate the treaty effect by focusing on its impact *on signatories*, we find that it is on average .04 larger for countries that restricted their current account in the

TABLE 1. Estimated Treaty Effects

	Selection Model, ρ_s and $\rho_n = 0$		Selection Model, ρ_s and ρ_n Vary	
Years Since Last Restriction	Mean	95% Confidence Interval	Mean	95% Confidence Interval
0	.137	(.095,.183)	.098	(.059,.142)
1	.164	(.026,.308)	.129	(−.035,.285)
2	.023	(−.042,.093)	.012	(−.057,.088)
3	.021	(−.036,.081)	.012	(−.049,.077)
4	.020	(−.029,.072)	.011	(−.042,.067)
5	.018	(−.025,.065)	.011	(−.036,.059)
6	.017	(−.021,.059)	.010	(−.031,.054)
7	.016	(−.019,.055)	.010	(−.027,.049)

Note: von Stein makes several modifications to the model presented by Simmons (2000), and yet attributes the entire difference between her estimates and Simmons' estimates to selection bias. By fixing ρ_s and ρ_n at zero and then estimating the effect of signing Article VIII using the model presented by von Stein, we can observe how much impact selection bias—as opposed to other changes in how the effect is modeled—impact the estimates. The left columns present the estimated impact of Article VIII in a case where we have imposed the requirement of no selection bias; the right columns are our replication on von Stein's estimates. Making this direct comparison while holding other modeling decisions constant, we see that all else equal, selection bias has only a minor impact on our estimate of the treaty effect. Accounting for selection bias reduces the estimated effect from .137 to .098 in the first year since the last restriction, and from .164 to .129 in the second year.

past year than the results she reports.[7] There is nothing wrong with the choice to report on this relationship, but for comparative purposes, she is not reporting estimates on exactly the same causal relationship as that reported in Simmons 2000. For those who might wish to implement von Stein's model in the future, it is critical to specify *a priori* precisely the causal relationship in which they are most interested.

Another reason we cannot attribute the full difference in estimates to selection effects is that von Stein simultaneously switches from a logit to a probit functional form. This decision alone reduces the mean estimated Article VIII impact by .04 for countries that last restricted 1 year ago. Using the logit model, we estimate the marginal effect of signing Article VIII when a country is 1 year from its last restriction as .26, with a 95% confidence interval from .20 to .32. When switching to a probit model, however, the estimated mean marginal effect drops to .22, with a 95% confidence interval from .17 to .27. Again, we have no problem with this choice, and we recognize the probit is necessary to generate her specific selection model. Our point is simply that her results are driven to some extent by making different distributional assumptions, and not by the "problem" of selection bias.

The most substantial difference between von Stein's estimate and Simmons's (2000) estimate comes from how they deal with time.[8] If Simmons had accounted for time using two dummy variables in the same way that von Stein does (for 0 and for 1 year since last restriction), the original article would have reported results that differ by only .007 from von Stein's for countries that had restricted in the prior year.[9] Simmons used splines instead (Beck, Katz, and Tucker 1998), a reasonable choice but not the only possible one. We are not accusing von Stein of handling the time dependence of observations inappropriately. But if Simmons had used two dummies in the original article, von Stein would not have had much of a case for a research note based on selection bias. Model dependencies, not selection bias, account for much of the gap between Simmons 2000 and von Stein.

Another way to show that selection bias may not account for the difference in results is by estimating separate probit models predicting restrictions for signatories and non-signatories. This is equivalent to estimating von Stein's selection model while fixing ρ_s and ρ_n both equal to zero. If von Stein is right that selection bias explains the majority of the change in the estimated effect, fixing ρ_s and ρ_n should lead the estimated Article VIII impact to return to something near its original estimate as presented in Simmons (2000). But as the similarity of the two estimates presented in Table 1 illustrates, that is far from the case. Imposing

[7] The marginal effect of the treaty on the mean nonsignatory is .10, with a 95% confidence interval from .06 to .14. For the mean signatory, though, the mean marginal effect increases to .14, with a 95% confidence interval from .08 to .19.

[8] As this note is primary concerned about selection effects, we do not enter into an extended discussion about how correctly to handle the time series issues. The original 2000 article utilized a set of two cubic splines, generated using Beck, Katz, and Tucker's BTSCS program for STATA. von Stein chose to use two dummy variables to control for time rather than for splines. Beck, Katz, and Tucker (1998) note that there should not be any substantive difference on the estimate coefficients, and they mention that they have a slight preference for using splines over the dummies.

[9] Consider countries that are 1 year past their last restriction. The updated version of Simmons's model estimates the marginal effect of being an Article VIII signatory as .124, with 95% confidence intervals running from .08 to .17. The selection model predicts a highly similar marginal effect of .129, where the 95% confidence interval runs from −.04 to .29. The two estimates differ, clearly, in their uncertainty. But they provide nearly identical estimates of the mean Article VIII impact, a result that should cause us to be cautious in concluding that selection bias is what accounts for the differences between Simmons and von Stein.

the condition of no selection bias, we observe estimated treaty impacts that are only slightly larger than those reported by von Stein. Hence other modeling decisions, including the switch from a logit to a two-stage probit model and to the use of dummy variables to control for time, must explain most of the reduction in the estimated treaty effect.

Even if we accept the smallest available estimates of the impact of Article VIII—those that come when Simmons's (2000) original model is modified to control for time using only two dummy variables instead of splines—two points are worth stressing. First, we dispute von Stein's conclusion that "a legal commitment to Article VIII appears to have little constraining power." Using the *smallest* estimates presented so far, 1 year after having restricted its current account, a country is likely to revert to restrictions with a probability of .24 if it is a nonsignatory, but just .11 if it has signed Article VIII. That is a change of over 50% in the probability of restricting, and it has proven quite robust—in not one of the specifications cited previously does the impact of the treaty become consistently statistically insignificant. Second, a convincing case has not been made that *selection bias* is what accounts for the bulk of the discrepancy between von Stein's results and those reported in Simmons. This becomes apparent only when we conduct a controlled methodological experiment and change one assumption at a time.

CONTROLLING FOR BIAS: PROPENSITY SCORE MATCHING

Our theory of how and why international law works implies that governments do not enter into legal commitments randomly. If they did, commitments would hardly be credible and markets would have no reason to take Article VIII commitments seriously. The theory suggests that treaties screen *and* constraint. Von Stein's estimator does not convincingly show that the screening effects *overwhelm* the constraining effects of legal commitments. Nor does her statistical model advance our understanding of the factors that lead governments simultaneously to commit and to comply with their legal obligations, because the bias is attributed partly to "unobservables." But we do accept that the inherent problem of selection bias is potentially very real and must be addressed. Only by doing so will skeptics warm up to the idea that treaties not only screen but also constrain governments' future behavior.

We advocate the following. *Commitment* should be modeled by using the event history style of analysis employed in the original 2000 article. Every effort should be made to theorize and to include in the commitment model all observables theory suggests are relevant, and an effort should be made to theorize and measure purported "unobservables" as well. And to estimate the treaty's *effect* on subsequent behavior, we advocate matching techniques informed by both theory and by the analysis of the decision to commit to the treaty. Nonparametric approaches such as matching control for bias on observables without making the strong distributional assumptions required by Heckman-type models. And in recent work, they have demonstrated their utility when confronting thorny problems related to nonrandom assignment to treatment as well. (Harding 2003; Imai, 2005).

Our point of divergence from von Stein is our contention that important influences on commitment and compliance can be theorized, observed, and (imperfectly) measured. The most reasonable "unobservable" for which we agree it would be desirable to control is a government's *political will* to remove restrictions from the current account. If a government truly is determined to liberalize its economy, then we should be able to find traces of this in policy areas distinct from but related to the current account. We should expect a government that is intent on a program of liberalization—independent from its Article VIII commitment—to implement other policies designed to liberalize trade and to encourage the freer international movement of capital. A number of observable measures of political will can be used in this context. We use three. First, a government that has opened up its economy to capital flows likely has the "political will" to become integrated into the world economy. Second, a government that has become a member of the General Agreement on Tariffs and Trade (GATT, which evolved in 1995 into the World Trade Organization, or WTO) is also likely to have some "political will" to liberalize. And finally, a government that is more democratic might pursue economic openness and eschew restrictions that deny free access to foreign exchange.[10] Democracy was included in the original (Simmons 2000) model of commitment, and found not to be a strong influence. When we reran the original model, we found that countries that had opened their capital account were highly likely (hazard ratio = 5.33; $p = .007$); GATT/WTO members were possibly likely (hazard ratio = 2.05; $p = .24$); and democracies were less likely a positive influence on Article VIII adoption (hazard ratio = 1.05; $p = .38$). A case can be made that these measures for political will should be taken into consideration when trying to determine the effect of Article VIII on the probability of restricting the current account.

In this section, we report the effects of Article VIII on restriction behavior estimated after a preprocessing matching step (Ho et al. 2004). Matching prior to performing standard parametric analyses reduces or eliminates the bias caused by selection on observable characteristics.[11] It also helps reduce the model dependency of our estimated effects, which is especially important given the sensitivity of the estimated effects to modeling decisions that we illustrated earlier. Using matching prior to implementing variants of the parametric model in Simmons 2000, we recover estimated treaty effects—defined as the average treatment effect—that are large and robust to model specification. For instance, the average treatment effect

[10] The measure used is the difference between democracy scores and autocracy scores, Polity IV dataset.

[11] For more on the theoretical foundations of matching, see Abadie and Imbens 2004, Imbens 2004, Rosenbaum and Rubin 1984.

for countries in the year after signing year is a reduction of 24.2 percentage points, with a 95% confidence interval from 3.9 percentage points to 43.1 percentage points. That is substantially closer to the estimated treaty effects of Simmons (2000) than those of von Stein. Simply put, Article VIII signatories are more likely to comply than are nonsignatories that are identical across a wide range of observed variables, including variables designed to proxy for political will.

To perform the matching, we first redefined our unit of observation to be a 6-year period of time during which we observe a country, or a "country-period." The 66 treated observations are countries that were Article VIII signatories for the first time in the fifth year of the observation window.[12] This allows us to observe the countries for 4 years prior to signing, for the signing year, and for 1 year following the signing year.[13] The universe of possible control cases includes all 6-year country-periods that do not overlap with the treated observations, for a total of 1,634 potential control cases. For instance, if Algeria never signs over the period of the dataset, but is observed for 30 years, it offers 25 possible control cases, one for each continuous 6-year period. And if Bangladesh signs in 1995, but was observed for 22 years before signing, the 13 prior periods that do not overlap with the treated observation period might provide potential control cases in those early years. This redefinition of the unit of observation would seem to markedly reduce the amount of data available to the researcher, but in fact it more accurately captures the transitions that we actually wish to observe—as well as the time-dependent structure of the observations.

Because this redefinition relies on combining 6 years of data, the data become far more sensitive to listwise deletion as a missing data strategy. We decided, then, to impute the missing covariates rather than discard the entire unit of observation (King et al. 2001) whenever data were missing. To do so, we used the mix package (Schafer 2003) in R 1.9.1. As a result, we have not one but five datasets, hence, five sets of matched observations. Estimating our causal effects across the five separate datasets allows us to incorporate the uncertainty that results from the imputation.

We then estimated a propensity score for each country-period in the new dataset using MatchIt (Ho et al. 2004b). A propensity score is the conditional probability of receiving the treatment (Rosenbaum and Rubin 1984)—that is, signing on to Article VIII after the fourth year—given the observed covariates. Doing so, we find that concerns about selection on *observables* are well justified: the mean propensity score for the signatories is approximately .43, whereas for the control cases, it is just .02. Most of the cases in our prospective control group have a very low conditional probability of receiving the treatment. In other words, they are simply not comparable to the treated cases. They are highly unlikely to sign onto Article VIII during the observation window, and thus not very useful in estimating treatment effects.

To identify well-matched control cases from the 1,634 candidates, we then followed the guidelines proposed by Ho et al. (2004a). We matched exactly on the single most important predictor, the average number of years of restrictions placed in the 4-year pretreatment observation period, and also matched on the estimated propensity score to achieve approximate balance on other covariates.[14] Of the 66 treated cases, we were able to match between 42 and 47 depending on the imputed sample. Initially, we tested for balance by ensuring that there were no significant differences in the treated and control samples on any of the 19 important covariates.[15] There were none. We looked for imbalanced samples by comparing all the possible multiplicative interactions of the 19 variables across the five matched datasets, and found just five significant differences out of the 1,805 possibilities. We then generated a list of other potentially unbalanced covariates by running sample *t*-tests and also the more powerful bootstrap Kolmogorov–Smirnov test[16] using the "Matching" package (Sekhon 2005) on all available Year 1 and Year 4 covariates. Any covariate whose *p*-value was under .10 on either test for any of the five matched datasets was noted. In all, the samples proved quite well balanced, with just 3 to 10 of the 37 covariates unbalanced for a given matched dataset. Not only do we have balance on most of the important predictors used by Simmons (2000) and von Stein, but we also have balance on our new measures of "political will" (capital account openness and GATT/WTO membership). Any estimated treatment effects, then, should be less vulnerable to concerns about political will as an omitted variable. And before running any parametric models, we have already identified those remaining confounders that could threaten our inferences.

[12] The dataset covers countries through 1997 and includes 100 transitions into signatory status, but only 66 countries are observed across the 6-year observation window.

[13] Both von Stein's work and our own suggests that these 6 years are the most crucial window to isolate treaty effects. von Stein illustrates that beginning at roughly 4 years before signing, restriction behavior begins to change, and both von Stein and Simmons (2000) find that the strongest impact of Article VIII is in the first few years after last placing a restriction. Here, we measure the effects as they vary across a unit of time that is different and perhaps more intuitive. Rather than looking at the effects by when countries placed their last restriction, we look at the effects as the time from signing increases.

[14] Even after matching, the treated and control groups differ in their propensity scores, so we used a caliper of .25 standard deviations to ensure that treated observations were not being matched to very different control-group observations.

[15] Here, "significantly different" is defined as any case where the *t* statistic comparing the means of the matched treatment and control groups is greater than 2. The "important covariates" are those that proved useful in estimating the propensity score. For Year 4 of the observation window—the year just prior to signing—these include restriction behavior, reserve volatility, reserves as a fraction of GDP, terms of trade volatility, GDP growth, GDP per capita, IMF surveillance, regional noncompliance, the use of IMF credits, the country's democracy score, its capital account openness, its status as a GATT/WTO member, and the calendar year. Also included in this test were the years of restriction-free behavior prior to the first year of the observation window, Year 1 gross national product (GNP) per capita, the number of years from joining the IMF to the beginning of the observation window, the use of IMF credits in Year 1, and the propensity score.

[16] For more on this test and its application to matched data, see Abadie 2002, Sekhon and Diamond 2005, and Sekhon 2004.

TABLE 2. P-Values for *t*-Tests and Bootstrap Kolmogorov-Smirnov Tests

	Mean, Treated	Mean, Control	*t* Test *p*-Value	KS Test *p*-Value
Year 1 Covariates				
Restrictions	0.705	0.727	0.816	
Years Since Last Restriction	2.159	2.455	0.793	0.976
Flexibility	1.386	1.455	0.523	
GNP/Capita	1976.245	1312.445	0.211	0.28
Change in GDP	3.663	3.925	0.896	0.604
Reserves/GDP	0.154	0.093	0.168	0.798
Reserve Volatility	−3.19	−3.209	0.906	0.758
Year	1987.023	1987.909	0.52	0.769
Terms of Trade Volatility	3.046	3.18	0.341	0.167
Universal	42.675	43.504	0.4	0.763
Regional Restrictions	34.574	27.698	0.275	0.062
IMF Surveillance	1.864	1.909	0.507	
Openness	78.148	82.117	0.758	0.058
GATT/WTO Member	1.705	1.705	1	
Balance of Payments/GDP	−4.353	−4.187	0.922	0.925
Use of Fund Credits	1.818	1.795	0.79	
Years of IMF Membership	25.159	24.795	0.887	0.498
Year 4 Covariates				
Restrictions	0.659	0.659	1	
Years Since Last Restriction	2.864	3.273	0.761	0.742
Flexibility	1.409	1.614	0.056	
GNP per Capita	2369.863	1958.404	0.491	0.916
Change in GDP	4.843	5.781	0.698	0.597
Reserves per GDP	0.172	0.159	0.781	0.301
Reserve Volatility	−3.186	−3.234	0.771	0.763
Terms of Trade Volatility	2.991	3.142	0.279	0.113
Universal	48.485	49.482	0.607	0.574
Regional Restrictions	42.14	36.61	0.361	0.092
IMF Surveillance	1.909	1.932	0.698	
Openness	79.596	81.073	0.908	0.308
GATT/WTO Member	1.75	1.773	0.805	
Balance of Payments/GDP	−4.444	−4.001	0.856	0.636
Use of Fund Credits	1.386	1.386	1	
Democracy	2.609	1.21	0.334	0.561
Capital Account Openness	1.134	1.114	0.772	0.945
Other Covariates				
Average Years of Restrictions, Years 1—4	0.705	0.705	1	1
Restrictions, Year 2	0.727	0.705	0.816	
Restrictions, Year 3	0.727	0.727	1	
Propensity Score	0.328	0.32	0.87	1

Note: This table presents *p*-values for *t*-tests and bootstrap Kolmogorov-Smirnov tests for the first of our five matched datasets. As is evident, the matched dataset is balanced on a wide range of covariates. Defining balance as all *p*-values higher than .05, these samples are balanced. Using our stricter criterion of $p > .10$, only four covariates are potentially unbalanced. And having identified those potential confounders, we can include them as explanatory variables in standard parametric models.

As an example of the balance obtained, consider the first of the five matches: across the 4 years prior to the potential treatment, the treated group and control group are identical in terms of their restriction behavior in Years 3 and 4, and differ only very slightly and insignificantly in Years 1 and 2. Or consider the calendar years of the matched groups. For the treatment group, the average year when the observation window begins is 1987, whereas for the control group it is 1988. As Table 2 makes clear, similarly close matches hold for the vast majority of variables in both Years 1 and 4 of the observation window. And of the 361 possible multiplicative interactions among the 19 key covariates, not one has a *t* score greater than 2. For the matched countries that sign onto Article VIII within the observation window, we have identified control cases that are indistinguishable across a wide range of measures, from the presence or absence of IMF surveillance to the countries' GNP per capita. In terms of all of the variables we have observed in the pretreatment phase, the matched pairs differ chiefly in that the treated group actually signed on to Article VIII in the fifth year, whereas the control group did not sign on during the observation window.

What, then, is the estimated effect of signing Article VIII? Here, we used probit models to calculate the average treatment effect for both the year of signing and the year after signing. Starting with Simmons's (2000) original model, we discarded those explanatory variables that were no longer approaching significance as predictors and tested those variables that were potentially unbalanced in any of the five matched samples as described earlier (in this case, Year 4 flexibility, Year 4 openness, and Year 1 regional restrictions). In the end, we wound up with a model of restriction behavior that closely approximated that presented by Simmons, with measures of openness in Year 1, democracy in Year 4, GATT/WTO status in Year 4, reserve volatility in Year 4, and an indicator variable marking whether the country last restricted in the previous year.[17] The estimated treaty effect in the signing year is a reduction of 17.7 percentage points, with a 95% confidence interval that runs from −0.7 percent points to 35.6 percentage points. For the year after signing, the same model leads to an estimated effect of 24.2 percentage points, with a 95% confidence interval from 3.9 percentage points to 43.1 percentage points. And as Ho et al. (2004) argue, using matching to preprocess the data reduces model dependency, and so should provide readers with added confidence that these results are not very sensitive to the specification of the parametric model, a point that our own data analyses confirm. These estimates are quite similar to those reported by Simmons, and confirm yet again that these estimated treaty effects are quite robust: they show up consistently across a wide range of modeling approaches and specifications.

To be sure, von Stein's critique is about nonrandom assignment to treatment owing to both observable and unobservable selection factors, and matching assumes that there is no selection on unobserved covariates. Certainly, though, matching can play a role in narrowing the range of possible unobservables, just as we demonstrated earlier. And it can also help in another way, even though it does not quantify the degree of selection on unobservables. By providing us with paired lists of countries that are highly similar on the observed covariates, matching allows us to draw upon knowledge not quantified within the data.[18] Instead of assuming that latent variables are distributed in a bivariate normal fashion—which is certainly not an assumption that researchers can observe in practice—we are making the more tangible assumption that each of our treated countries is similar in all important respects to its matched control. If there is indeed some unobservable influence, whether it is political will or anything else, careful study of the paired list in combination with substantive knowledge of the cases will help us understand what it might be. This is precisely what we advocate scholars do. We think this approach will yield far more insights into the selection effects in making international legal and other kinds of commitments than will fragile statistical methods that allow theoretically interesting processes to go unobserved.

[17] That is, we adopt the strategy for time dependence employed by von Stein, although we drop the second indicator variable, as it is zero across all observations.

[18] Ho et al. direct readers to (Rosenbaum 2002) chapter 3 for more on this point.

REFERENCES

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *Journal of the American Statistical Association* 97: 284–92.

Abadie, Alberto, and Guido Imbens. 2004. "Large Sample Properties of Matching Estimators for Average Treatment Effects." http://emlab.berkeley.edu/users/imbens/sme_21 jan 04.pdf (July 12, 2005).

Albert, A., and J. A. Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrick* 71 (1): 1–10.

Beck, N., J. N. Katz, and R. Tucker. 1998. Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science* 42 (4): 1260–88.

Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43 (4): 1209–30.

Christmann, Andreas, and Peter J. Rousseeuw. 2001. "Measuring Overlap in Logistic Regression." *Computational Statistics and Data Analysis* 37: 65–75.

Downs, George W., David M. Rocke, and Peter N. Barsoom. 1996. "Is the good new about compliance good news about cooperation?" *International Organization* 50 (3): 379–406.

Harding, David J. 2003. "Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out and Teenage Pregnancy." *American Journal of Sociology* 109 (3): 676–719.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator of Such Models." *The Annals of Economic and Social Measurement* 5: 475–92.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–98.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2004a. "Matching as Nonparametric Preprocessing for Improving Parametric Causal Inference." http://gking.harvard.edu/files/abs/matchup-abs.shtml (January 12, 2005).

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2004b. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference, Version 1.0-1." Harvard University, Cambridge, MA. Computer program available at http://gking.harvard.edu/matchit/.

Imai, Kosuke. 2005. Do Get-Out-The-Vote Calls Reduce Turnout?: The Importance of Statistical Methods for Field Experiments. *American Political Science Review* 99 (2): 283–300.

Imai, Kosuke, Gary King, and Olivia Lau. 2005. "Zelig: Everyone's Statistical Software, Version 2.2-2." Harvard University, Cambridge, MA. Computer program available at http://gking.harvard.edu/zelig/.

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86 (1): 4–30.

King, Gary, James Honaker, Anne Joseph, and Kenneth F. Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (1): 49–69.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 41–55.

Lee, Lung-fei. 1984. "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity." *Econometrica* 52 (4): 843–63.

Lee, Lung-fei. 2001. "Self-Selection." In *A Companion to Theoretical Econometrics*, ed. B. H. Baltagi. Malden, MA: Blackwell Publishers Inc.

Lemke, Douglas, and William Reed. 2001. "War and Rivalry Among Great Powers." *American Journal of Political Science* 45 (2): 457–69.

Liao, Tim Futing. 1995. "The Nonrandom Selection of Don't Knows in Binary and Ordinal Responses: Corrections with the Bivariate Probit Model with Sample Selection." *Quality and Quantity* 29: 87–110.

R Development Core Team. 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R_project.org.

Reed, William. 2000. "A Unified Statistical Model of Conflict Onset and Escalation." *American Journal of Political Science* 44 (1): 84–93.

Rosenbaum, Paul E., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–24.

Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. New York, NY: Springer Verlag.

Rousseeuw, Peter J., and Andreas Christmann. 2004. "The Noverlap Package," version 1.0–1. Available online at http://cran.r-project.org/doc/packages/noverlap.pdf.

Sartori, Anne E. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11: 111–38.

Schafer, Joseph. 2003. "The Mix Package: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data." http://cran.r-project.org/src/contrib/Descriptions/mix.html (January 14, 2005).

Sekhon, Jasjeet S. 2004. "The Varying Role of Voter Information across Democratic Societies." http://jsekhon.fas.harvard.edu/papers/SekhonInformation.pdf (February 24, 2005).

Sekhon, Jasjeet S. 2005 "Matching: Multivariate and Propensity Score Matching with Automated Balance Search." Computer program available at http://jsekhon.fas.harvard.edu/matching.

Sekhon, Jasjeet, and Alexis Diamond. "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies." Presented at the Summer Meeting of the Society for Political Methodology, Tallahassee, FL.

Signorino, Curt. 2003. "Structure and Uncertainty in Discrete Choice Models." *Political Analysis* 11 (4): 316–44.

Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs. *American Political Science Review* 94 (4): 819–35.

Timpone, Richard J. 1998. "Structure, Behavior, and Voter Turnout in the United States." *American Political Science Review* 92 (1): 145–58.

Vreeland, James Raymond. 2003. *The IMF and Economic Development*. New York: Cambridge University Press.

Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18: 327–50.

Winship, Christopher, and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25: 659–707.