Do Nonproliferation Agreements Constrain?*

Bradley C. Smith† William Spaniel‡

September 28, 2020

Abstract

One way nuclear agreements might keep signatories from proliferating is by constraining nuclear capacity. Theoretical work on nonproliferation often points to such constraints as an important driver of nonproliferation success. Some have argued that, absent sufficient constraint, states with the desire and capability to proliferate will do so. Faced with more costly routes to a weapon, states subject to technological constraint may abide by the terms of the deal. This perspective poses an important empirical question: do nonproliferation agreements result in significant technological constraint in practice? This paper evaluates the empirical prevalence of constraints arising from nonproliferation deals. Doing so requires (1) providing an appropriate measure of nuclear proficiency and (2) developing an estimate of the counterfactual, no-agreement capacity of states that received such agreements. This study addresses both of these points. First, new data is gathered to estimate proficiency, improving upon existing measures in the literature. Second, the generalized synthetic control method is applied to estimate counterfactual proficiency levels for the recipients of agreements. With this approach, the constraining effects of deals the United States implemented with Japan, South Korea, and Taiwan and the Declaration of Iguaçu between Brazil and Argentina is evaluated. The findings indicate that the constraining effect of these nonproliferation agreements is minimal.

*Conditionally accepted, *Journal of Peace Research*

†(Corresponding Author) Department of Political Science, Vanderbilt University (bradley.carl.smith@gmail.com)

‡Department of Political Science, University of Pittsburgh.

1 Introduction

In 1969, President Richard Nixon announced the Guam Doctrine, indicating that allies were responsible for their own security. South Korea responded by accelerating its nuclear program. Washington began worrying that another state would soon join the nuclear club. Like with most other nonproliferation negotiations, the United States engaged with South Korea on many dimensions (Drezner, 1999; Solingen, 2007). One pathway was to help develop the Kori-2 power plant, with the expectation that Seoul would halt objectionable nuclear activities.

Provisions like this within nonproliferation agreements—deals designed to make a party less likely to develop nuclear weapons—have a similar goal. Washington sought to constrain South Korea's capacity to produce nuclear weapons. With less research done in nuclear science, racing to the bomb became harder. Such constraint has been cited as important for nonproliferation generally, and researchers identify constraints as an important nonproliferation mechanism (Reiss, 1995; Mehta, 2020).

Nevertheless, researchers have yet to rigorously evaluate whether states actually constrain in a meaningful way. Doing so requires operationalizing capacity in a fine-grained manner and identifying a proper counterfactual, neither of which is straightforward. In this paper, we overcome these problems by expanding a dataset to estimate nuclear capacity and creating synthetic counterfactual states. For the cases we can analyze, deals lead to little constraint. At most, states may reduce their capacity by a small degree relative to where they would have been without a deal.

To preview the method, we begin by strengthening Smith & Spaniel's (2020) nuclear capacity measure. Compared to the existing measure, our estimates capture considerably more nuance in changes to a state's abilities and contains substantially less noise. We then bring the generalized synthetic control (GSC) method to the topic (Xu, 2017). This allows us to estimate a counterfactual version of the state that had not entered the agreement. Comparing the counterfactual state's capacity to the observed level provides an estimate of the agreement's effect.

We apply the method to four agreements. Three are U.S. security guarantees to East Asian countries: Japan, South Korea, and Taiwan. We find minimal support for the notion of a proficiency-constraining effect. The agreements between allied East Asian countries would seem to be a likely case to observe a substantial effect. Japan, South Korea, and Taiwan had warm relations with the U.S. and therefore had less concern that policymakers in Washington would backtrack on a deal. Nevertheless, we observe continued capacity increases following the agreements. They fall below the estimated counterfactual versions of these countries that did not broker a deal. However, the overall effect is substantively minor and statistically indistinguishable from zero. Individually, only Taiwan shows some partial restraint.

The fourth is the 1985 Declaration of Iguaçu, a bilateral agreement between Argentina and Brazil. Given the relative ineffectiveness of the East Asian agreements, one would expect to see little restraint between Argentina and Brazil. After all, these countries were rivals at the time of their agreement. The results confirm that speculation. Neither country increased their capacity following the agreement. However, we estimate the counterfactual countries would not have improved much either. Thus, the estimated effect is close to zero.

These subtleties demonstrate the utility of the GSC method, as some observations may suggest the wrong inference. For example, because Argentina's and Brazil's proficiency scores plateau after 1985, a simple comparison before and after would suggest that the deal cooled both states' appetite for development. The GSC findings indicate no effect. Meanwhile, South Korea's score continues to increase after its deal. This would seem to imply that the agreement might have encouraged South Korea to continue. Yet the GSC method finds at best marginal improvement over the counterfactual.

For clarity, our findings speak to a state's general ability to produce nuclear technology, not any one specific technology. One goal of such agreements is to convince would-be proliferators to not develop enrichment and reprocessing technology. Qualitative analysis of agreements suggest that this is effective. But production of those technologies and the ability to produce them are distinct concepts (Smith & Spaniel, 2020). Furthermore, the Non-Proliferation Treaty gives states access to nuclear technology, and assistance correlates with proliferation behaviors (Fuhrmann, 2009). We estimate that signing agreements has little effect on overall capacity; the estimates are consistent with a state's nuclear industry not disappearing following a deal but rather shifting to uncontroversial activities. Still, in some ways, this gives reason for optimism: states are not pursuing those forbidden technologies despite the ability to do so. But it underscores how technical incapacity does *not* drive that reticence.

2 Observed Constraint and Nonproliferation Mechanisms

Nonproliferation agreements are common.[1] They range from global agreements like the Non-Proliferation Treaty to smaller initiatives like the Iran Deal and the Agreed Framework. Sometimes, the primary signatories are allies; sometimes they are antagonists. In each case, these agreements aim to alter the decision calculus of a potential proliferator. One important factor in this calculus that agreements often target is a state's technical capacity to develop a weapon. Our goal is to assess whether this effect exists empirically: do nonproliferation agreements constrain a country's technical capacity? The question we pose is empirical, but the answer we provide has important implications for identifying which theoretical mechanisms underlie nonproliferation success.

Empirically assessing whether agreements lead to constraint is important because an existing theoretical perspective points to constraint itself as the primary determinant of nonproliferation success. According to this perspective, constraining technology is a *necessary* condition for long-run nonproliferation compliance. Absent constraints on technology, threatened states with the capability to develop a weapon will do so (Thayer, 1995). Consequently, nonproliferation agreements *must* result in the dismantling or halting of infrastructure development to work over the long-term.

This perspective is consistent with findings from the supply-side literature that identify technical and industrial capability as a strong predictor of proliferation (Singh & Way, 2004; Jo & Gartzke, 2007). Similarly, countries that receive more nuclear assistance proliferate more frequently (Fuhrmann, 2009; Brown & Kaplow, 2014). Put simply, a country that lacks cost-effective production capacity will not proliferate.

In this view, technical capacity constraint is key to assuring nonproliferation success. The means through which this constraint may arise are varied. Under threat of preventive war or sanctions, states may roll back existing technical infrastructure or abandon planned advances. Strategic concessions may also create incentives for states to limit technical capability (Spaniel, 2019; Mehta, 2020). The end result is the same: technical capacity is constrained as the result of a nonproliferation deal. Under the perspective outlined above, constraint is the key component. In its absence, agreements have little hope for long-term success.

[1]See Bas & Coe (2018: 613) for a thorough list.

Emphasis on technical constraint at least appears to be part of the nonproliferation regime. The U.S. classifies fuel cycle development as a proliferation risk. Building nuclear weapons requires fissile material. A state can obtain that from enriching uranium or reprocessing spent unenriched or low-enriched uranium to gather fissile plutonium. Controlling the nuclear fuel cycle therefore controls a state's ability to weaponize.

Consequently, limiting this ability and monitoring existing facilities is a major part of the nonproliferation regime. This shows in many agreements. The JCPOA requires that Iran close multiple facilities, end ongoing construction, hand over machinery, and downblend enriched uranium. Libya gave the U.S. much of its nuclear infrastructure in its 2003 agreement. The Agreed Framework called for North Korea to halt its activities in exchange for less-concerning light water power plants. Heading toward its agreement with Argentina, Brazil's 1988 constitution banned non-peaceful nuclear applications. Argentina never opened a plutonium reprocessing facility it started constructing in 1978, and closed a uranium enrichment facility in 1994. More generally, a goal for Atoms for Peace is to convince nuclear scientists worldwide to pursue peaceful technology.

Another perspective suggests that technical constraint, though beneficial, is not directly necessary for nonproliferation success. This perspective points to the manipulation of strategic incentives as the key mechanism. Rather than requiring technical constraint as a necessary condition, success hinges on a more general alteration of a proliferator's strategic incentives. Technical capacity plays a role in these strategic considerations, and these manipulations often aim to affect such capacity. However, unlike the previous perspective, work that points to these broader strategic incentives as the key mechanism indicates that proliferation can be prevented *even if a state maintains a high level of technical capacity.* The key difference is whether technical constraint is a necessary condition for preventing proliferation. Under the first perspective, states follow through on nonproliferation agreements because they lack the technical capacity to break them. This alternative view suggests that, as a result of positive and negative inducements, states *choose* not to break agreements because doing so is unattractive.

Work in this vein indicates a variety of inducements used to alter incentives and prevent proliferation. On the positive side, conciliatory policies reduce risks of proliferation (Paul, 2000; Debs & Monteiro, 2016; Bas & Coe, 2018; Spaniel, 2019; Mehta, 2020). Opponents offer the time-adjusted benefits of nuclear weapons. By doing so, proliferation is no longer worthwhile. On the negative side, breaking an agreement risks economic conflict and international tensions. For example, sanctions raise the price of developing weapons (Solingen, 2007; Miller, 2014). Meanwhile, the credible threat of preventive war also causes states to reevaluate proliferation decisions (Debs & Monteiro, 2016).

These nudges need not be permanent to be useful. Simply buying time can convince a current hawkish leader to pause nuclear weapons production. Concessions may also empower doves within the potential proliferator to extend their tenure in office. Alternatively, from the 'demand side' perspective (Hymans, 2006; Solingen, 2007), additional time gives more opportunity to resolve underlying tensions, thereby overriding the long-term proliferation need.

The logic of inducement suggests that agreements can work even in the absence of significant capacity reduction. Though agreements aim to limit some activities, they often leave key capacities intact. Why? Coercion is the central issue. Agreements must be commensurate with the state's ability to proliferate. Making nuclear weapons more expensive therefore incentivizes opponents to reduce the concessions granted. It also means that the potential proliferator must endure the hardships of the negative inducements for longer. As such, even if would-be proliferators refrain from prohibited activities, they may still maintain or improve their overall capacity. For example, if the skill learned in civilian projects compensates for the lost experience on sensitive technologies, then the constraining effect is illusory. All that changes is what scientists are *currently* working on, not what they *can* achieve.

Researchers recognize this connection. Nuclear 'hedgers' retain the capacity to build a nuclear weapon but stop short. The threat to build allows states to obtain 'some deterrent benefits without paying the costs of overt proliferation' (Narang, 2017: 120). Levite (2003: 72) further argues that the 'leverage it provides [reinforces] a state's coercive diplomacy strategy.' As such, reticent would-be proliferators evade nuclear controls (Gheorghe, 2019). Opponents recognize this. A core component of the nonproliferation regime is technology sharing. In fact, it is possible that technology sharing enhances a country's proficiency (Fuhrmann, 2009; Brown & Kaplow, 2014).

These perspectives gives rise to a variety of empirical expectations. The first perspective suggests that agreements should result in constrained technical capacity. The second perspective is also potentially consistent with observed constraint. However, the second perspective is also consistent with a lack of constraint, as the hedging and bargaining literatures suggest. At present, the theoretical perspectives offered by the literature are consistent with either possibility: a constraining effect of agreements, or a lack of one. These possibilities point to an important empirical question: do nonproliferation agreements lead to technological constraint? We state each of these possibilities as hypotheses below before outlining what can be learned from the answer.

Hypothesis 1. *After reaching an agreement, a potential proliferator's capacity to build nuclear weapons is lower compared to the counterfactual world where they did not reach an agreement.*

Hypothesis 2. *After reaching an agreement, a potential proliferator's capacity to build nuclear weapons minimally changes compared to the counterfactual world where they did not reach an agreement.*

How do these hypotheses relate to the theoretical perspectives outlined above? First, it is important to note that the mechanisms outlined above are not mutually exclusive. Many deals exhibit elements of both. Indeed, positive and negative inducements often result in states limiting their technical capacity. The key difference in the theoretical perspectives is whether nonproliferation *requires* technical constraint. According to the first perspective, the answer is yes: potential proliferators with the technical capability to build the bomb will do so. Under the second perspective, even highly technically capable states forego proliferation when the inducements are right. This perspective is consistent with either hypothesis—agreements may constrain, but they may succeed even without imposing significant constraints.

This difference illuminates the stakes of our empirical analysis. If we were to find evidence consistent with the first hypothesis, this would be consistent with *either* theoretical perspective. However, we find evidence consistent with the second hypothesis. The absence of constraint indicates that there is little empirical evidence that the first mechanism is operating. Failure to find an overall constraining effect provides indirect evidence that positive and negative inducements are doing the heavy lifting.

3 Empirical Challenges

Two barriers stand in the way of estimating whether nuclear agreements constrain a country's technical capacity. Fortunately, recent methodological developments provide solutions. We now explain those challenges and how we address them.

3.1 Measurement Problems

The first major challenge is in measuring the dependent variable. In the quantitative nuclear weapons literature, binary proliferation behaviors tend to receive most of the focus. From a measurement perspective, binary variables lack nuance. From an inference perspective, they give us little leverage to explain why states reach those outcomes. This is of particular importance for the cases we are interested in. Nonproliferation agreements the U.S. has reached with allies, for example, see consistent compliance. Explaining that success requires analysis of a state's nuclear program beyond its binary proliferation status.

Smith & Spaniel (2020) provide some leverage here, developing a continuous measure of proficiency by estimating a standard three-parameter item response model. This model takes as inputs observable indicators of proficiency and produces an estimate of a state's overall nuclear proficiency. By incorporating multiple measures and weighting them appropriately to generate a single proficiency estimate, this approach builds upon additive indices (Jo & Gartzke, 2007), which treat all activities as equally informative.[2]

Using this technique, Smith & Spaniel (2020) generate a pair of measures to assess different aspects of the proliferation process. One of these measures, ν Infrastructure, divorces nuclear science capability from nuclear weapons capability. It is a useful measure for a country's price tag for nuclear weapons; the higher a state's infrastructure, the fewer scientific hurdles the country needs to overcome. It therefore appears to be an ideal variable for testing how agreements affect a state's proficiency.

However, we have more work to do. Although the IRT method is well-suited for the construction of the variable, their actual measurement is crude, using only 12 activities. Two problems result. First, a country's score can fluctuate wildly upon beginning or quitting an activity. It is unclear whether a country's proficiency actually spikes or plummets in these years or if it is just an artifact of the minimal data. Second, the ν measure has large uncertainty intervals.

[2] To describe the technique more formally, let i index states, t index years, and k index indicators. The item-response model posits that observed activity k for state i in year t denoted, y_{it}^k, occurs according to the process $Pr[y_{it}^k = 1] = logit^{-1}[\beta_k(\nu_{it} - \alpha_k)]$. The parameters β_k, ν_{it}, and α are then estimated via standard Bayesian techniques, with ν_{it} representing the desired proficiency estimate.

To rectify this problem, we have collected additional data. The supplemental appendix gives a full accounting of the expansion. The data now cover previously ignored or under-exploited activities. As Smith & Spaniel (2020) discuss, the variation in the type of new activities we include is an asset; each reveals new information about a state's underlying proficiency.

To begin, we revisited the U.N. Statistical Yearbooks to disaggregate the industrial activities that Meyer (1986) originally coded and Jo & Gartzke (2007) condensed. The original data just had indicators for heavy water and non-heavy water power plants; we now include an indicator for reactors with each type of neutron moderator. The updated measure also captures whether a country uses the plant for desalination, if it meets a high production threshold, advanced power plant designs, and membership in working groups. We also include information on countries' research reactors using the IAEA Research Reactor Database. Disaggregating Fuhrmann & Tkach (2015) uranium enrichment data, we separate whether states have advanced centrifuge technology. We track facility types from the IAEA's Nuclear Fuel Cycle Information System (NFCIS), which Herzog (2020) identifies as critical to a state's nonproliferation calculus. In our final major upgrade, we scanned the International Nuclear Information System database for countries publishing nuclear research. The dataset now contains 104 variables in total, including 98 that fall under the infrastructure category. We also expand the data to add 15 years more coverage.

Two improvements to the data make it particularly useful for our purposes. First, we code whether a state engages in key activities on their own or with international partners. This is critical given that some nuclear agreements result in a transfer of nuclear assistance. Second, the data better plots the trajectory of the facilities. This means providing separate codings for facilities under construction and the scale of those facilities. The scale allows us to better track a state's ability to acquire fissile material to construct a weapon, which deals often target. For example, the JCPOA called for Iran's Arak facility to have its reactor filled with concrete and its Fordow Fuel-Enrichment plant to be converted to a nuclear science center.[3]

[3]One component missing from the dataset is a state's quantity of fissile material. Accurate codings of this are impossible to acquire because they are state secrets. However, we do not believe this is a significant absence for what we want to measure. While fissile material is necessary to produce a nuclear weapon, actual possession of it only addresses a proximate problem, not the underlying problem. A state with both enriched uranium and production facilities does not lose the ability to proliferate after giving up its enriched uranium because production facilities can replace it. The extensive coding of

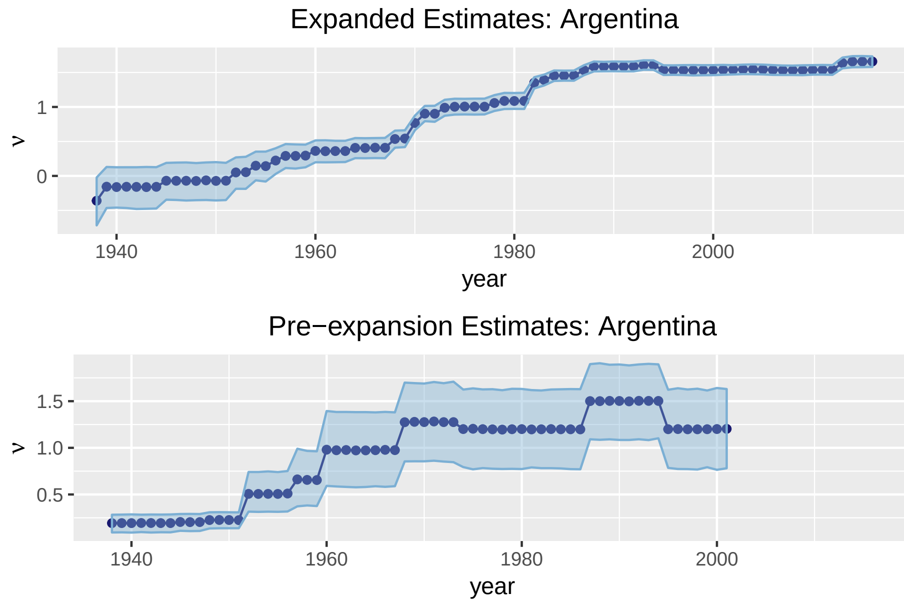

Figure 1. A comparison of our new estimates with Smith & Spaniel's (2020) original estimates.

Figure 1 illustrates the improvement, plotting Argentina's estimated proficiency with the original and our new measure. Two points stand out. First, the broad trends in the data are similar, with Argentina's proficiency generally increasing over time. In fact, the ν measures for all countries have a correlation of .905. Second, the new measure has fewer wild fluctuations. For example, the original formulation sees Argentina's capacity plateau from 1987 to 1994. This coincides with Argentina's opening of the Pilcaniyeu Enrichment Plant. Our new measure marginally moves during that time.

The new data also improves the measure in other important ways. For example, Sagan (2011) critiques Jo & Gartzke's (2007) data used to construct Smith & Spaniel's (2020) original measure. Sagan points to two cases in which these data produce substantively unappealing predictions. First, under the Jo & Gartzke measure, Trinidad and Tobago have a higher level of capability than North Korea in 2001. Second, the Jo & Gartzke measure assigned Egypt with a higher capability score than South Africa in the same year (Sagan, 2011: 329–330). To test the validity of the new measurements, we investigated these cases. While the original Smith & Spaniel measure and the Jo

Fuhrmann & Tkach (2015) and the NFCIS allow us to address the underlying problem, which we think is the critical element to capture a meaningful change in a state's nuclear capability.

& Gartzke measure assigned similar rankings, our estimates alleviate Sagan's (2011) concerns. We present these improvements in the appendix.

More broadly, the estimates are more precise than the original analysis. The new data have more activities observed, providing more information and reducing the corresponding uncertainty around the estimates. We can see this by comparing the credible intervals generated in each set of estimates. To do this, we calculate the ratio, for every country-year, of the absolute value of the ν estimate to the length of the corresponding 95% credible interval obtained from the posterior distribution. The metric jumped from about 0.91 to 2.59, a substantial reduction in uncertainty. Replications of Jo & Gartzke (2007) and Fuhrmann & Tkach (2015) also showed that the new measure generates a better model fit.

3.2 Control Problems

Our second issue is the fundamental problem of causal inference. To understand how a deal changes proficiency, we need to compare an agreeing party to itself, except if it had not made the agreement. This is impossible—we cannot observe the counterfactual outcome directly. The solution relies on leveraging differences between states that did and states that did not receive such agreements. As usual, the key difficulty here is that both treatment assignment and observed outcomes may be related to unobserved confounders, rendering direct comparisons unreliable.

One solution is to assume that all such unobservable factors are unchanging over time. If so, inclusion of country-specific dummy variables eliminates the confounding influence. This approach, commonly referred to as fixed-effects regression, is commonplace in the quantitative literature on nuclear proliferation. However, the key assumption necessary for this strategy to uncover a causal effect is problematic in our setting.

To illustrate, consider a major trend in nuclear proliferation: the spread of nuclear knowledge over time. To demonstrate that time-varying, country-specific factors likely drive some of the variation in our data, we estimate a simple regression of proficiency on time. If the relationship is not constant, then the assumption necessary for fixed effects estimation is questionable. Table I gives the results. The estimated coefficient is positive and statistically distinguishable from zero, indicating an upward trend. The

TABLE I. OLS regression assessing the relationship between time and capacity.

	Dependent variable:
	Smith & Spaniel's (2020) ν Score
Time	0.012^{*}
	(0.0004)
Constant	-23.440^{*}
	(0.861)
Observations	10,986
R^2	0.063
F Statistic	740.411^{*} (df = 1; 10984)

Note: $^{*}p<0.01$

simple fixed-effects approach is not suitable.

When the assumption of time-invariant heterogeneity is implausible, a 'differences-in-differences' strategy provides a possible alternative. This approach allows for time-varying confounders by invoking the *parallel trends* assumption, which supposes that the difference between pre-and post-treatment outcomes for the treated and control groups are the same in the absence of treatment. Put simply, parallel trends requires that the *trajectory* of unobserved confounders is identical

The parallel trends assumption is also problematic in our setting. Though untestable, it is more plausible when observed pre-treatment trends appear parallel in the data. This is not the case for proficiency—some states rush to develop proficiency, while others pursue a more leisurely path. The Soviet Union, for example, sacrificed efficiency for speed, worrying that it would face preventive measures from the U.S. (Spaniel, 2015). In contrast, India took a leisured approach, testing its first weapon in 1974 but not producing a deliverable weapon until 1987. Differences such as these suggest that the parallel trends assumption is implausible. While trends are not parallel in general, a possible solution is to select a plausible 'control group' by hand, selecting states that appear to have parallel development trajectories in the pre-treatment period to our treated countries of interest. While this is possible in principle, it relies on arbitrary choices about which states constitute acceptable controls. Consequently, any results obtained from such an approach risk being driven by these choices.

4 Creating Counterfactual Potential Proliferators

Summarizing the previous section, the standard tools employed to wrest causal estimates from time-series data are ill-equipped to overcome concerns about unobserved confounding here. Fortunately, Xu (2017) provides a solution. The generalized synthetic control (GSC) method directly addresses the difficulties with fixed effects and differences-in-differences designs. Rather than relying on unrealistic assumptions about the nature of unobserved confounding variables, the GSC method estimates the influence of unobservables in a flexible, semiparametric manner. As a result, the GSC uses a weaker identification assumption than a standard pooled regression, allowing for time-variant, unobserved heterogeneities across units (Xu, 2017: 62–64). Below we provide a brief overview of the method; we refer readers to Xu (2017) for the technical details.

The GSC approach has two steps. The first step leverages information in the control group to address concerns with unobserved confounders. Doing so generates estimates of two quantities: time-varying factors and country-specific factor loadings. The time-varying factors account for unobserved trends, such as nuclear technology diffusion or changes in the overall security environment. As such, the time-varying factors are roughly analogous to flexible time trends. The country-specific factor loadings determine how these time-varying factors influence each specific country. For example, we might think of technology diffusion as an unobservable, time-varying factor. But technology diffusion does not affect all states in the same way. Country-specific factor loadings weight the influence of each state's unobserved factors, allowing unobservable diffusion to influence some states more than others.

The second step leverages the estimates of unobserved factors to generate predicted values of the counterfactual outcomes for the states that received a deal. Using the factor estimates from the first step, the method generates factor loadings for the treated observations to maximize fit between predicted and observed outcomes in the pre-treatment period. Next, this model is used to generate predictions in the post-treatment period. These predicted values function as a 'synthetic control'—one may interpret them as estimates of a recipient state's technological progress had they not formed a nonproliferation agreement. The deal's estimated causal effect is the observed outcome compared to the estimated counterfactual from the synthetic control.

Importantly, there is a connection between the GSC approach and the typical approaches outlined above;[4] the GSC captures difference-in-differences and fixed effects models as special cases (Xu, 2017). Additionally, while the spirit of the approach is similar to Abadie, Diamond & Hainmueller's (2015) synthetic control method, the GSC approach has a number of advantages that make it particularly well-suited for our application.[5] The method allows us to deal with multiple treated groups in a single run and produces frequentist uncertainty estimates. This allows us to apply Rainey's (2014) guidance for assessing null findings. Further, the GSC removes the need to select the appropriate pre-treatment model specification by using a cross-validation procedure to select the appropriate model for unobserved factors. Finally, because the GSC calculates the counterfactual using regression techniques, it is close to existing quantitative work on nuclear proliferation (Singh & Way, 2004; Jo & Gartzke, 2007; Fuhrmann, 2009; Brown & Kaplow, 2014).

Although this method is well-suited to deal with potential unobserved factors driving proliferation behavior, we must account for observed confounders. Indeed, the key identifying assumption of the GSC is that all relevant observable confounders are incorporated.[6] In practice, this requires us to include as controls all observables that are correlated with both the probability of a nonproliferation agreement and a recipient's proficiency. The controls we include are in line with those included in existing work in the literature (Singh & Way, 2004; Jo & Gartzke, 2007; Fuhrmann, 2009; Brown & Kaplow, 2014). To account for a state's relevant security environment, we control for the number of recent disputes, the number of borders, and major power status. To account for the confounding effect of a nuclear ally, we include dummies for US and Soviet Union/Russia alliance. To control for a state's ability to commit to agreements, we account for regime type and NPT ratification. Finally, we control for economic

[4]More precisely, the method captures unobserved factors that take the form of common shocks with potentially heterogeneous cross-sectional effects. Formally, allowing i to index country and t to index time, the GSC can account for an unobserved factor U_{it} as long as the unobserved factor can be decomposed into a multiplicative form $U_{it} = a_i * z_t$.

[5]One notable drawback is that, unlike the Abadie, Diamond & Hainmueller (2015), the GSC does not provide the 'weights' used to construct the synthetic counterfactual. This is because the GSC has a regression framework, and thus the weights are not the target of estimation. We view this as a tradeoff, and we select the GSC method because of the advantages described above.

[6]This is the 'strict exogenity' assumption, which requires that the error term is independent of treatment assignment, observed covariates, and unobserved cross-sectional and time-variant factors.

factors by including measures of GDP, imports, and exports.[7] The GSC model includes state fixed effects to account for time-invariant state-level factors.

Before proceeding to our case selection and analysis, it is also important to note that, though the GSC method improves on existing approaches in a number of ways, it is not a panacea. The GSC method shares some potential weaknesses with the prior approaches discussed above. In particular, fixed-effects, difference-in-differences, and GSC methods approaches implicitly rely on a unit homogeneity assumption. As King, Keohane & Verba argue, this is a fundamental assumption in all scientific research, both quantitative and qualitative (1994: 94). If potential proliferators are very different from all other states in ways that cannot be accounted for, this assumption may not hold in practice. Addressing this is an important topic for future quantitative work on proliferation. Though providing a solution to this potential issue is outside the scope of the present analysis, we flag it in order to clarify more precisely how our approach does improve on existing work.

5 Case Selection

Having developed the model to build counterfactual countries, we now explore which agreements to investigate. Unlike cross-country regression approaches that estimate the effect of treatment variables using pooled data, the GSC method delivers targeted estimates for smaller groups of states. Thus, our results come with an important caveat. Rather than aiming to estimate an average effect across a large number of recipients, we specify a smaller group of states that received agreements with similar features, then use the GSC method to generate the appropriate counterfactuals, and analyze the resulting estimated effect. Consequently, we emphasize internal over external validity, aiming to credibly assess the impact of a small number of agreements that are well-suited for our method. This approach aligns with Samii's (2016) call for targeted research designs that emphasize clarity and specificity over generality. Our estimates have an interpretation similar to those of a targeted case study that might use available qualitative data to make claims about counterfactual, no-agreement outcomes. Rather than doing this with qualitative evidence, the GSC method leverages quantitative data to provide such counterfactuals, bridging qualitative and quantitative approaches (Abadie, Diamond &

[7] See the appendix.

Hainmueller, 2015).

To choose which cases to study, we consulted Bas & Coe's (2018) list of proliferation agreements. The method then guided which to cover. GSC requires training data for a state prior to the treatment year. This eliminated the Soviet successor states, which declared independence in 1991. We also require a sufficient time period after the treatment year to obtain a more complete picture of the agreement's effect. Many of the control variables we use end in 2000. This precludes the 1994 Agreed Framework for North Korea.[8] It also eliminates Libya's 2003 agreement and the 2015 Iran Deal. Finally, we cannot analyze system-wide agreements, thereby disqualifying the NPT.

This leaves us with five cases to analyze. We now describe the intention of the agreements and how we can apply the treatment to the method.

5.1 East Asian American Allies

The U.S. reached three similar agreements with East Asian allies in the 1970s. Negotiations with Japan came first. Despite being the victim of the only nuclear attacks in history, China's test in 1964 and observations from the Vietnam War concerned Japanese conservatives (Paul, 2000: 48). Nevertheless, legal barriers stood in the way to a nuclear deterrent. Article 9 of the Japanese constitution outlaws war for dispute resolution. Meanwhile, 1955's Atomic Energy Basic Law restricts Japan to peaceful nuclear research. Yasuhiro Nakasone therefore commissioned a white paper to find a loophole. It concluded that tactical, defensive weapons would be legal (Campbell & Sunohara, 2004: 222).

For the U.S., Japan's nuclear push came at a bad time. The NPT opened for signature in 1968, and U.S. policymakers worried of a world twenty nuclear weapons states. Creation of a worldwide nonproliferation regime that outlawed nuclear weapons was the solution. Japanese nuclear weapons acquisition a few years into the movement represented a major setback.

The U.S. therefore looked to convince Japan to adhere to the developing nonproliferation norm. To do so, it placed a key bargaining chip into the pot: Okinawa. Washington had maintained control over the island after World War II. To begin the

[8] Nevertheless, we add North Korea to the analysis in the appendix for the years available. Our main results do not change much. North Korea observes a small drop in its proficiency relative to the synthetic control.

transfer process, President Richard Nixon signed the 'Agreed Minute.' Control of Okinawa would revert to Japan, though the U.S. could reacquire it during a crisis (Roehrig, 2017: 50). In 1972, the U.S. initiated the transfer. Since then, Japan mostly dropped its overt push.

South Korea had a similar experience. Following the Korean War, the U.S. maintained a large deployment of American soldiers. But with support for the Vietnam War fading, Nixon began a new approach. The U.S. began playing a secondary role in the defense of its allies, beginning with the announcement of the Guam Doctrine. Although the Guam Doctrine also affected Japan, policymakers in Seoul showed greater concern. Following through on his pledge, Nixon removed 20,000 soldiers from the Korean Peninsula (Hong, 2011: 487). President Park Chung Hee began searching for a solution. Without Washington, Park believed a nuclear deterrent would compensate.

By 1974, American policymakers had become suspicious. Washington escalated pressure against Seoul the following year and offered inducements, deploying 600 nuclear weapons to South Korea (Choi, 2014: 76) and $1.5 billion in military aid (Drezner, 1999: 255). It also threatened to withhold loans (Solingen, 2007). Standing firm put the construction of its Kori-2 nuclear power plant in jeopardy (Drezner, 1999: 260-261). Meanwhile, Canada stipulated that South Korea had to ratify the NPT as a precondition for receiving a CANDU reactor (Choi, 2014: 75-76).

The final case of U.S. intervention into East Asian allied nuclear pursuit is with Taiwan. China was Taiwan's major security concern. Following the Chinese Civil War, mainland China held a massive military advantage. Taiwan therefore pursued a military relationship with Washington. Formalized in 1954, the Sino-American Mutual Defense Treaty provided aid and military support in the event of an attack on the island.

Those terms were satisfactory until China's nuclear test, Nixon's warming to the mainland, and the Guam Doctrine. Taiwan began reconsidering its posture. It sought foreign assistance, purchasing reactor technologies from Western nations and South Africa. The American embassy recognized Taiwan's progress in 1973 after noticing incongruities between the technology transferred and Taiwan's civilian programs (Hersman & Peters, 2006: 544). One year later, the CIA believed that Taiwan possessed a small-scale weapons program (Albright & Gay, 1998: 57). Washington removed its foreign-deployed nuclear weapons from the island.

By 1977, the U.S. clarified its nonproliferation preference and articulated its demands. If Taiwan continued the program, it risked losing its current benefits (Hersman & Peters, 2006: 544). The U.S. already supplied the nuclear fuel necessary to run the island's nuclear power. It also could weaken its security guarantee. Aside from any potential intervention against China, Washington could terminate arms sales. Although Taiwan's formal nuclear program lingered afterward, the threat marked the end of significant consideration.

In each of the cases, key policymakers and analysts have attributed the success to technical constraint. Indeed, Washington's global nonproliferation push came in part due to concerns for Japan's growing technical ability (Kase, 2001: 56). The level of restrictions placed on South Korean frustrated the latter's scientists (Pollack & Reiss, 2004: 259). And the U.S. exerted ample effort to monitor and stop technical progress in Taiwan (Mitchell, 2004: 298–300).

5.2 The Declaration of Iguaçu

Argentina and Brazil faced a period of heightened tensions from the 1970s. Neither was sure of the other's plans for the continent, and both had military governments. Consequently, each began building its nuclear infrastructure to hedge against the other (Reiss, 1995: 45-52).

Nevertheless, Argentina and Brazil built a path toward rapprochement. The major breakthrough came in 1985 when Argentinian President Raúl Alfonsín met Brazilian President José Sarney at Foz do Iguaçu. The summit aimed to reduce mistrust, work toward mutual economic development, and integrate each other's civilian nuclear programs to reduce fears of a military escalation. Before leaving, the presidents codified these goals in the 'Declaration of Iguaçu.'

Whereas earlier attempts at an agreement had faltered, the Declaration of Iguaçu made steady progress. The leaders made annual trips to each other's countries. Nuclear scientists started visiting restricted facilities (Reiss, 1995: 55). Trade policies improved and the states enjoyed warmer relations. A firmer agreement came in 1991 with the Guadalajara Accords, which committed both parties' nuclear energy programs to peaceful applications and established a bilateral inspection regime. Brazil made positive steps by closing its Cachimbo test site in 1990. Meanwhile, Argentina ended its uranium enrichment in 1994. Like with the East Asian cases, policymakers and analysts viewed the technical constraint as critical to the bilateral agreement's success.

5.3 Applying the Treatment

These sets of cases share similarities but also have differences. For the East Asian countries, an ally provided the primary inducement to shift nuclear policy. The concessions were also given with the implicit goal of nonproliferation. And the exchanges were bilateral. With Argentina and Brazil, the rival states engaged with each other directly. The agreement was explicitly nuclear.

As a result, we must choose how to define the treatment variable. Given those differences, we adopt a conservative approach and split the analysis in two. One covers just the treatment of U.S. inducements in East Asia, and the other covers specifically the Declaration of Iguaçu. In each, we insert the other as a control for the estimated counterfactual.

We also must make a few choices about the year of the treatment. The above discussions explain our primary choices of 1973 for Japan, 1976 for South Korea, 1978 for Taiwan, and 1986 for Argentina and Brazil as the first full post-treatment years. A reasonable reading of some of these cases may yield a different treatment year. We defer these dates to the appendix as robustness checks.

6 Synthetic Control Results

With the method and cases of interest described, we now present the results. For each of the agreements, we first present individual effects, graphing comparisons of the observed and counterfactual development paths for each recipient state. Visualizing the individual effects provides an initial check on the validity of the results, allowing us to assess whether the observed and counterfactual levels of capability in the pre-treatment period are appropriately close. The state-level visualizations also allow us to detect possible heterogeneity in the effects that would be lost in the estimated average effects.

After presenting the state-level results, we consider average effects. Fortunately, the GSC method provides yearly estimates of the average effect. This means that we can assess not only whether these agreements slow or encourage development, but also whether their effect grows or diminishes as time passes from the agreement date.

6.1 East Asian Agreements

Our estimates for Japan, South Korea, and Taiwan appear in Figures 2, 3, and 4. In each graph, the solid line plots the observed capacity level, while the dashed line indicates the GSC's estimated counterfactual. The vertical line denotes the year of agreement. To avoid extrapolating too far, we focus on a ten-year window afterward. For each recipient, it is reassuring that the observed and counterfactual levels of proficiency in the pre-treatment period are very close. This suggests that the estimated counterfactuals fit the data generating process.

Focusing on Japan's development, Figure 2 clarifies the utility of the GSC approach. Looking only at the observed path of Japan's nuclear development, the rate of growth slows in the years after the agreement. Based solely on this information, one might conclude that the agreement caused this plateau in Japan's capacity. However, our estimates indicate that this is not the case. The estimated counterfactual also plateaus following 1972. In fact, the point estimate exceeds the counterfactual within a few years. This indicates that the agreement did little to slow Japan. For South Korea, we see a similar plateauing effect in the years following the 1975 agreement. However, our counterfactual estimate reveals an interesting wrinkle: this plateau only occurs in the aftermath of an *acceleration* of nuclear development. In the long-term, U.S. forward deployments and military aid appear to have reduced South Korea's capability by a modest amount. Taiwan also follows a distinct pattern. In the immediate aftermath of the treatment, the observed and counterfactual paths are close. After five years, however, they diverge as Taiwan's development flattens.

While the individual graphs are informative with respect to each case, we have not addressed a crucial question: did these agreements limit average capacity development? Viewing the individual effects simultaneously suggests that these deals did little to constrain capability in the short term. In each case, the counterfactual estimate lies close to the observed capability level in the first five years. For South Korea, the observed level even rises above the counterfactual, suggesting that the deal may have contributed to a modest short-term *increase* in South Korea's capability. The longer-term effects of these deals is murkier. For Japan, the deal seems to have made little difference, as

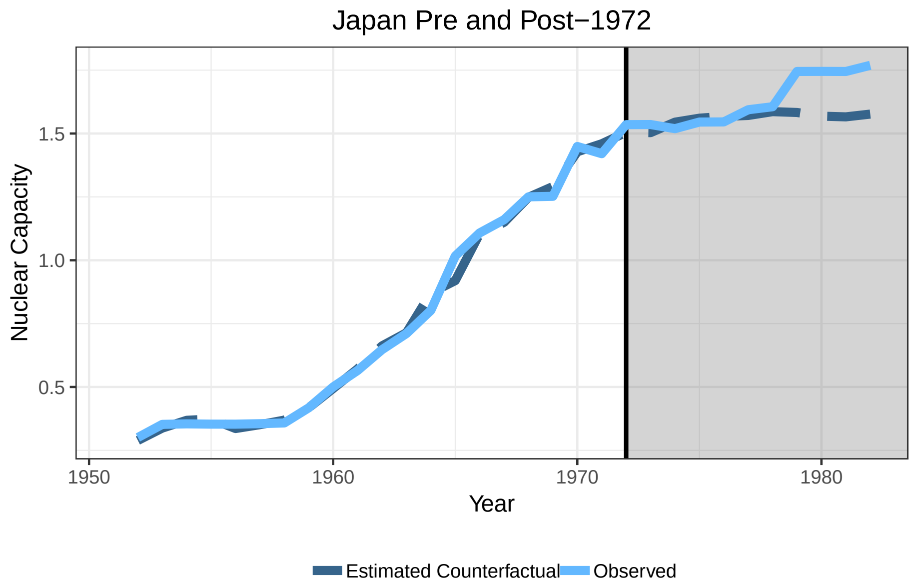

Figure 2. Observed and estimated capacity for Japan.

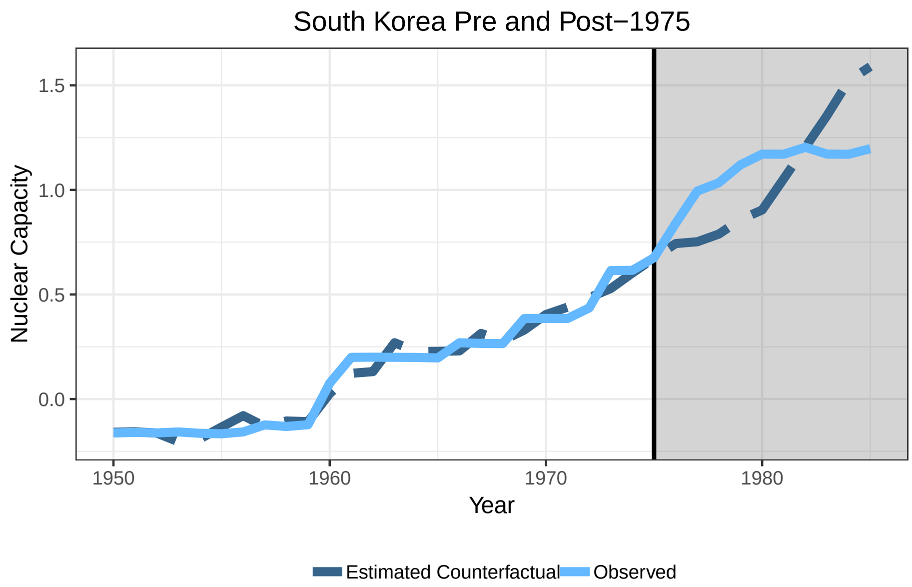

Figure 3. Observed and estimated capacity for South Korea.

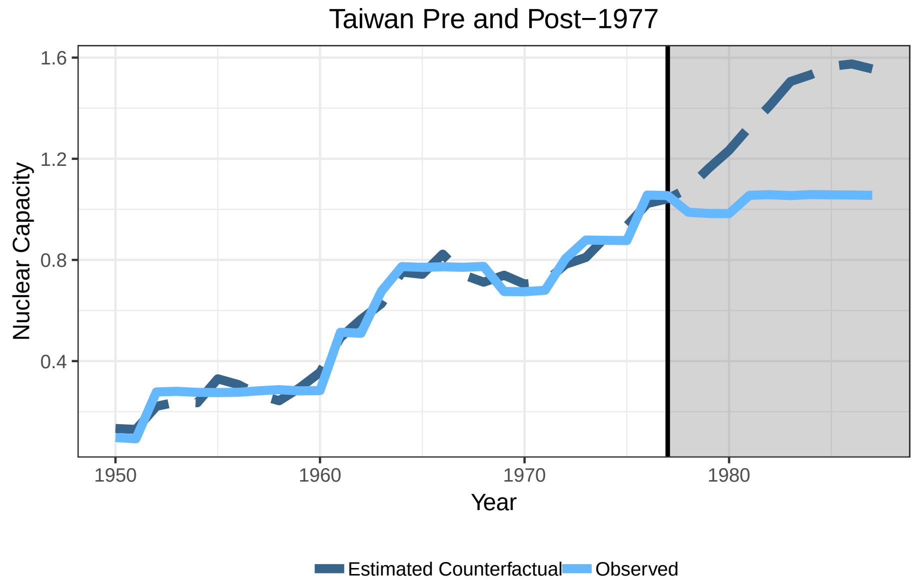

Figure 4. Observed and estimated capacity for Taiwan.

our counterfactual estimate deviates little from observed capability. However, the point estimates are suggestive of moderate long-term constraining effects for South Korea and Taiwan.

Of course, we can only conclude so much by considering point estimates of individual treatment effects. They tell us nothing about the average impact, nor do they reveal information about statistical precision. Fortunately, the GSC method provides us with estimates of the average effect and a relevant confidence interval. Moreover, the GSC provides *yearly* estimates of this effect. This allows us to draw conclusions about both the short and long-term impact of the agreements.

Figure 5 visualizes the average treatment on the treated (ATT) estimates for the East Asian deals. Estimated average effects for each year in the decade post-deal are presented, along with a 95% confidence interval. The point estimates of the ATT in each year align with expectations drawn from the individual plots: the effects are small, but grow through the post-agreement decade. More telling are the confidence intervals. Strikingly, in every year, we fail to reject the possibility that these agreements have no effect on the development of a recipient's capability.

While the estimates clearly indicate that we cannot reject a null effect in each year of a post-agreement decade, this finding is only partially satisfactory. Failure to reject

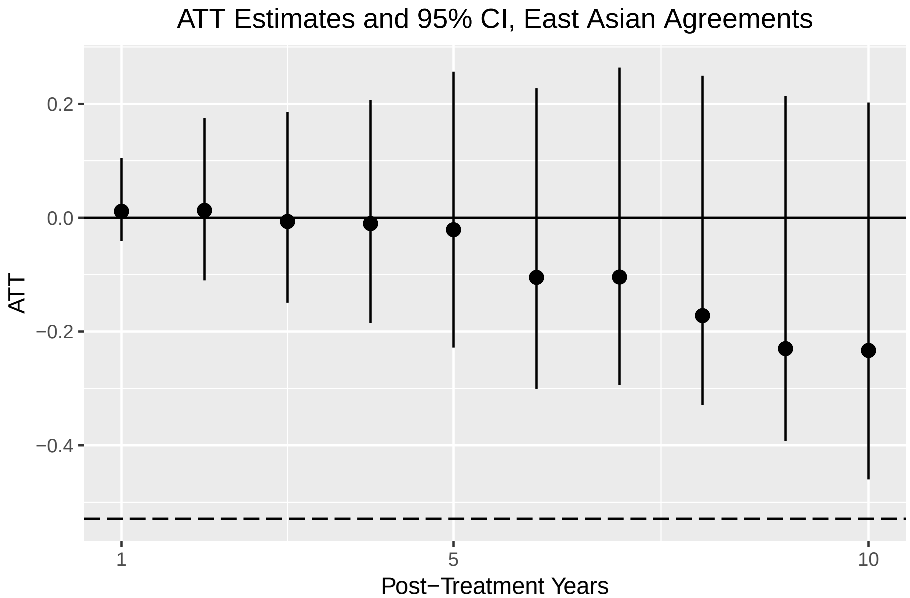

Figure 5. Yearly ATT estimates for East Asian agreements.

the null hypothesis does not necessarily constitute evidence of a substantively negligible effect (Rainey, 2014). Indeed, because the confidence intervals include zero, it is possible that the data are consistent with large positive or negative effects (Westlake, 1979). Put more directly: a confidence interval may include zero, but it may also include large positive or negative values.

Given this, how should we assess the range of effects that our estimates are consistent with? Rainey (2014) provides guidelines for researchers to reject substantively large effects. The approach is straightforward: researchers first define a 'meaningful' effect, then check whether such effects lie within the bounds of a 90% confidence interval.[9] If the confidence interval does not include the defined effect size, then we can confidently conclude that the data are not consistent with large effects.

How do we define a substantively meaningful effect? We let the data guide us here. Using the updated capacity scores, we gather the capability of all states that have tested nuclear weapons as of 2016. From this subset, we calculate the average difference in each state's ν score at the time they began to pursue nuclear weapons to their ν score at the time of their first successful test. We then take the average of these differences.

[9]We present a more conservative 95% interval above. The results are substantively identical with a 90% interval.

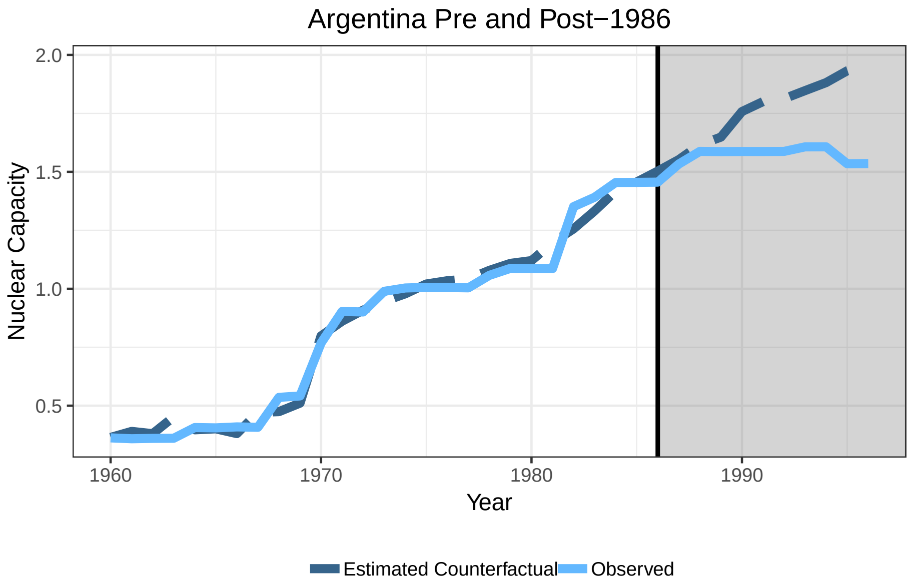

Figure 6. Observed and estimated capacity for Argentina.

The resulting number is approximately 0.53, which we can interpret as the average increase in a state's capability from the time it begins pursuing a weapon to the time of its first test. This amount gives us an idea of how much technological constraint is necessary, on average, to revert a state to its pre-pursuit level. We use this value as a benchmark for substantive significance.

Given that benchmark, we return to Figure 5. The dotted horizontal line is the benchmark. Following the guidelines above, for the first nine years after an agreement, we can reject a large effect. In fact, for many years we can reject substantially smaller effects. The interval does not include an effect half as large as the threshold until six years post-agreement. Only after a decade can we not reject a sizable effect, as the confidence interval barely overlaps the threshold at that point. This means that the largest possible effect consistent with the data after a decade barely meets the average level of development among states that have sought and obtained a weapon. We take this as evidence that the impact of these deals on the development of recipient capability in the first decade after an agreement is relatively small.

Summarizing, our analysis of the East Asian agreements suggests two preliminary conclusions. First, the influence of these agreements is statistically indistinguishable from zero. Second, while our estimates grow less precise as the time after treatment

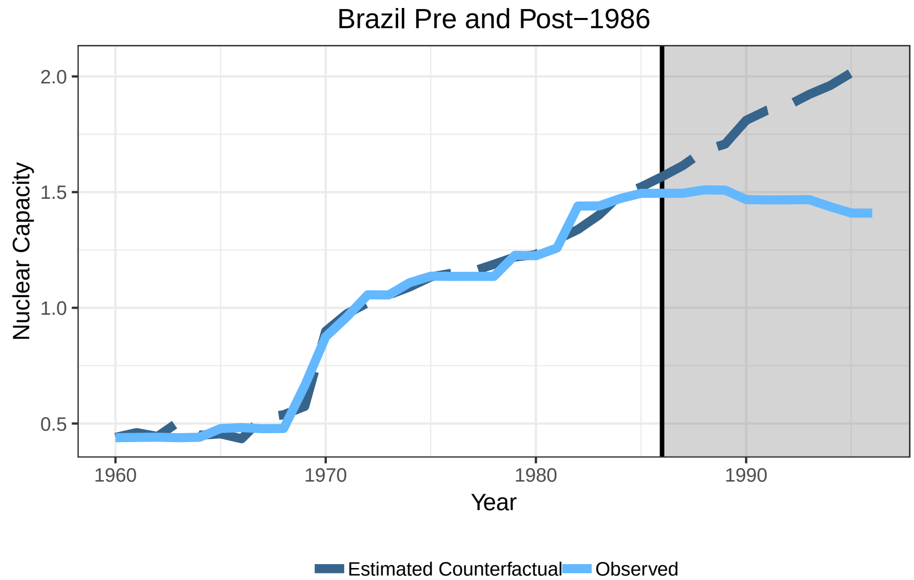

Figure 7. Observed and estimated capacity for Brazil.

grows, we can reject substantively meaningful effects throughout the decade after agreement.

6.2 The Declaration of Iguaçu

We plot the state-level observed and counterfactual paths of development in Figures 6 and 7. A first glance at the individual point estimates for Argentina and Brazil suggests a consistent pattern. In the first five years after the agreement, both lie slightly above their counterfactuals. Much like the East Asian allied agreements, the bilateral agreement between Brazil and Argentina seems upon first blush to have made little short-term difference. However, the long-run point estimates in each case deviate from the counterfactual.

Is this a significant effect? As with the East Asian agreements, we present yearly point estimates and confidence intervals for the ATT. A similar pattern emerges. While the estimated ATT grows over time, we fail to reject the null hypothesis of no effect in every year throughout the post-agreement decade. Particularly striking are the near-zero point estimates in the first four years. The point estimates only diverge from zero beginning in the fifth year, growing steadily for the remainder of the post-treatment

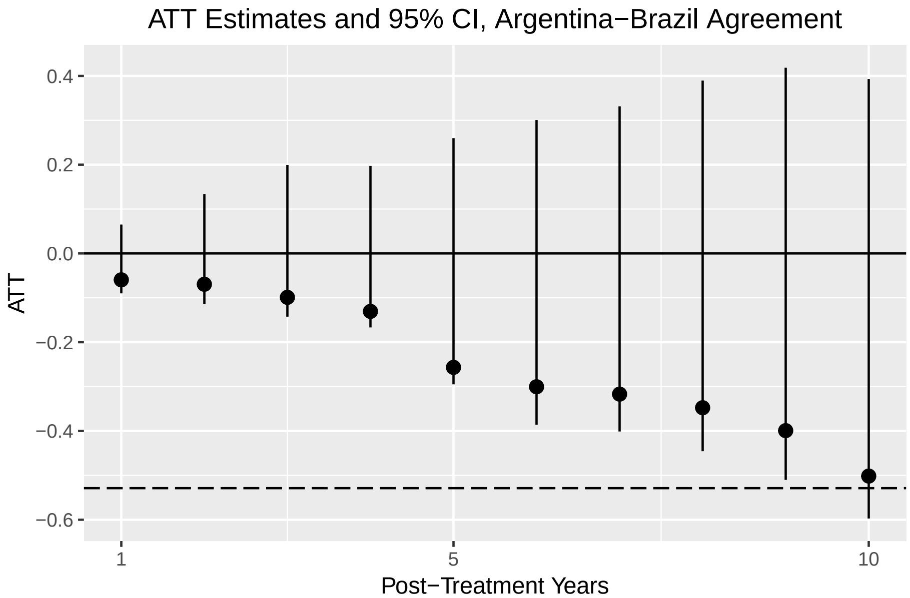

Figure 8. Yearly ATT estimates for Declaration of Iguaçu.

decade. While the point estimates do grow over time, they are never distinguishable from zero, as the confidence intervals indicate.

In addition to this, we perform the same exercise as before, assessing the degree to which we can rule out substantively large effects as statistically implausible. Similar to with the East Asian agreements, the confidence intervals indicate that in nearly every year in the post-agreement decade, the Declaration of Iguaçu is inconsistent with anything but a negligible effect. Only in the final year do the confidence intervals include a meaningful effect. Even then, the interval barely covers the threshold.

Taking stock of all the agreements, some patterns emerge. First, the effects are indistinguishable from zero. However, the pattern is more nuanced. We can more conclusively rule out large effects in the short-term. In particular, Figures 5 and 8 indicate that our estimates are precise within the first five years post-agreement. Within this window, we can rule out all but extremely small effects. Moving out of this window, our estimates grow less precise, but are still only consistent with relatively small substantive effects. Only towards the end of the post-treatment decade can we not convincingly reject the possibility of meaningful effects.

Finally, it is important to note that our inability to conclusively reject the possibility of meaningful effects in the long-term does not provide evidence that there is a substantively meaningful effect. While we cannot rule out a substantively meaningful effect ten years post-agreement, we also cannot rule out a null effect or a small positive effect. Our estimate of the long-term ATT is noisy. This is due to the GSC's accounting for the uncertainty about unobserved heterogeneity in the long-term. As the counterfactual estimates extrapolate further from the year of treatment, so too does uncertainty about these estimates.

6.3 Robustness

It is important to assess whether our results hold under reasonable alternative specification. These key robustness checks address the possible ambiguities in the date of treatment. We recode the year of treatment for Japan to 1971 to mark when the Japanese Diet adopted Three Non-Nuclear principles and Nixon's signature of the Agreed Minute. We recode the year of treatment for Taiwan to 1979 to account for the clarification of U.S. policy that accompanied the passage of the Taiwan Relations Act. While the informal agreement between Argentina and Brazil came in 1985, the more finalized Guadalajara Accords occurred in 1991. We therefore recode the year of treatment from 1985 to 1991. In each case, the new treatment date does not alter the substantive findings of our initial specification.

Taiwan's status provides another source of complication. The measure of capability that we adopt includes activities which Taiwan is precluded from participating in. We address this difficulty in two ways. First, we exclude Taiwan from the analysis, estimating the effect of U.S. agreements with Japan and South Korea. Second, we include Taiwan but reestimate our capability score, including all variables except for the IAEA research variables. Our substantive results remain unchanged.

While there are a number of issues with traditional approaches like fixed-effects regression, we performed a simple fixed-effects regression for comparison. The results are qualitatively similar to those from the GSC method. However, the fixed-effects results do not allow for us to analyze the year-by-year effects, and they are still plagued by concerns over time-varying unit-specific confounding.

7 Conclusion

We evaluated whether nuclear agreements constrain the development of nuclear technology. To address this question, we first sought out a suitable measure of technological development. We identified Smith & Spaniel's (2020) scores as appropriate and extended them by collecting additional data on activities indicative of nuclear development. This data generation exercise produced a measure suitable for evaluating the technological development processes we wish to study. Nevertheless, we still faced the difficulty of constructing the appropriate counterfactual. The GSC provides a desirable solution, and we therefore estimated the effect of nuclear agreements with it.

Our estimates indicate that agreements minimally constrain. Any reduction in proficiency is statistically indistinguishable from zero. Put differently, we cannot rule out the possibility that these agreements do not constrain recipient development of nuclear technology. Going a step farther, we assessed the uncertainty in our estimates, analyzing what range of effects we *can* rule out. In the first decade after an agreement, our estimates eliminate all but substantively small effects. Comparing to documented cases of proliferation, our estimates suggest that the maximum constraint plausible given our estimates would not stop proliferation on average.

The GSC method limited our analysis to those five cases. Our approach emphasized specificity over generality. We can only provide claims for the cases at hand. It is possible other countries—e.g., North Korea, Libya, or Iran—would produce similar results due to their distrust of the U.S. committing to long-term concessions. However, the U.S. may also believe these states are less likely to comply and thus would require greater constraints. As time passes, more data on these countries will become available. Researchers therefore may wish to revisit this question in the future.

These results have important implications for policymakers. If our results are right, constraint has little impact on long-term compliance with nonproliferation agreements. This suggests that other mechanisms are responsible for compliance. However, our study does not provide direct evidence of this, and researchers should continue exploring that part of the puzzle.

Finally, while we can confidently rule out meaningful constraining effects in the short and medium term, our estimates of the long-term impact of such agreements are noisier. Thus, if these deals do have any efficacy at constraining the development of nuclear technology, such effects only arise in the long term. This means that the success of efforts to constrain technology should not be judged based on short-term changes. We suggest caution in declaring deals that do not result in immediate reversal a failure.

References

Abadie, Alberto; Alexis Diamond & Jens Hainmueller (2015) Comparative politics and the synthetic control method. *American Journal of Political Science* 59(2): 495–510.

Albright, David & Corey Gay (1998) Taiwan: Nuclear nightmare averted. *Bulletin of the Atomic Scientists* 54(1): 54–60.

Bas, Muhammet A & Andrew J Coe (2018) Give peace a (second) chance: A theory of nonproliferation deals. *International Studies Quarterly* 62(3): 606–617.

Brown, Robert L & Jeffrey M Kaplow (2014) Talking peace, making weapons: Iaea technical cooperation and nuclear proliferation. *Journal of Conflict Resolution* 58(3): 402–428.

Campbell, Kurt M & Tsuyoshi Sunohara (2004) Japan. In: Kurt M Campbell; Robert J Einhorn & Mitchell B Reiss (eds) *The Nuclear Tipping Point*. Washington, DC: Brookings, 218–253.

Choi, Lyong (2014) The first nuclear crisis in the korean peninsula, 1975–76. *Cold War History* 14(1): 71–90.

Debs, Alexandre & Nuno P Monteiro (2016) *Nuclear politics: The strategic causes of proliferation*. New York: Cambridge University Press.

Drezner, Daniel W (1999) *The sanctions paradox*. New York: Cambridge University Press.

Fuhrmann, Matthew (2009) Spreading temptation: Proliferation and peaceful nuclear cooperation agreements. *International Security* 34(1): 7–41.

Fuhrmann, Matthew & Benjamin Tkach (2015) Almost nuclear: introducing the nuclear latency dataset. *Conflict Management and Peace Science* 32(4): 443–461.

Gheorghe, Eliza (2019) Proliferation and the logic of the nuclear market. *International Security* 43(4): 88–127.

Hersman, Rebecca KC & Robert Peters (2006) Nuclear u-turns: Learning from south korean and taiwanese rollback. *Nonproliferation Review* 13(3): 539–553.

Herzog, Stephen (2020) The nuclear fuel cycle and the proliferation danger zone Forthcoming, *Journal for Peace and Nuclear Disarmament*.

Hong, Sung Gul (2011) The search for deterrence. In: Byung-Kook Kim & Ezra F Vogel (eds) *The Park Chung Hee era*. Cambridge: Harvard University Press, 483–512.

Hymans, Jacques EC (2006) *The psychology of nuclear proliferation*. Cambridge: Cambridge University Press.

Jo, Dong-Joon & Erik Gartzke (2007) Determinants of nuclear weapons proliferation. *Journal of Conflict Resolution* 51(1): 167–194.

Kase, Yuri (2001) The costs and benefits of japan's nuclearization: an insight into the 1968/70 internal report. *The Nonproliferation Review* 8(2): 55–68.

King, Gary; Robert O Keohane & Sidney Verba (1994) *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Levite, Ariel E (2003) Never say never again: nuclear reversal revisited. *International Security* 27(3): 59–88.

Mehta, Rupal N. (2020) *Delaying Doomsday*. Oxford: Oxford University Press.

Meyer, Stephen M (1986) *The dynamics of nuclear proliferation*. Chicago: Chicago University Press.

Miller, Nicholas L (2014) The secret success of nonproliferation sanctions. *International Organization* 68(4): 913–944.

Mitchell, Derek J (2004) Taiwan. In: Kurt M Campbell; Robert J Einhorn & Mitchell B Reiss (eds) *The Nuclear Tipping Point*. Washington, DC: Brookings, 254–253.

Narang, Vipin (2017) Strategies of nuclear proliferation: How states pursue the bomb. *International Security* 41(3): 110–150.

Paul, TV (2000) *Power versus prudence*. Montreal: McGill-Queen's University Press.

Pollack, Jonathan D & Mitchell B Reiss (2004) South korea. In: Kurt M Campbell; Robert J Einhorn & Mitchell B Reiss (eds) *The Nuclear Tipping Point*. Washington, DC: Brookings, 254–253.

Rainey, Carlisle (2014) Arguing for a negligible effect. *American Journal of Political Science* 58(4): 1083–1091.

Reiss, Mitchell (1995) *Bridled ambition*. Baltimore: Johns Hopkins University Press.

Roehrig, Terence (2017) *Japan, South Korea, and the United States Nuclear Umbrella*. New York: Columbia University Press.

Sagan, Scott D (2011) The causes of nuclear weapons proliferation. *Annual Review of Political Science 14*: 225–244.

Samii, Cyrus (2016) Causal empiricism in quantitative research. *The Journal of Politics* 78(3): 941–955.

Singh, Sonali & Christopher R Way (2004) The correlates of nuclear proliferation: A quantitative test. *Journal of Conflict Resolution* 48(6): 859–885.

Smith, Bradley C & William Spaniel (2020) Introducing ν-clear: a latent variable approach to measuring nuclear proficiency. *Conflict Management and Peace Science* 37(2): 232–256.

Solingen, Etel (2007) *Nuclear logics: Contrasting paths in East Asia and the Middle East*. Princeton: Princeton University Press.

Spaniel, William (2015) Arms negotiation, war exhaustion, and the credibility of preventive war. *International Interactions* 41(5): 832–856.

Spaniel, William (2019) *Bargaining over the bomb*. New York: Cambridge University Press.

Thayer, Bradley A (1995) The causes of nuclear proliferation and the utility of the nuclear non-proliferation regime. *Security Studies* 4(3): 463–519.

Westlake, Wilfred J (1979) Statistical aspects of comparative bioavailability trials. *Biometrics* 35(1): 273–280.

Xu, Yiqing (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1): 57–76.