Introducing ν-CLEAR: A Latent Variable Approach to Measuring Nuclear Proficiency*

Bradley C. Smith† William Spaniel‡

November 10, 2017

Abstract

The causes and consequences of nuclear proficiency are central to important questions in international relations. At present, researchers tend to use observable characteristics as a proxy. However, aggregation is a problem: existing measures implicitly assume that each indicator is equally informative and that measurement error is not a concern. We overcome these issues by applying a statistical measurement model to directly estimate nuclear proficiency from observed indicators. The resulting estimates form a new dataset on nuclear proficiency which we call ν-CLEAR. We demonstrate that these estimates are consistent with known patterns of nuclear proficiency while also uncovering more nuance than existing measures. Additionally, we demonstrate how scholars can use these estimates to account for measurement error by revisiting existing results with our measure.

*Forthcoming, *Conflict Management and Peace Science*. We thank Andrew Coe, Matthew Fuhrmann, Sig Hecker, Daniel Hill, Brenton Kenkel, Roseanne McManus, Emily Ritter, Scott Sagan and the scientists at the Los Alamos National Laboratory for their comments on earlier drafts. We are also grateful to Eva Lin Feindt and Sharon George for their research assistance.

†Assistant Professor, Department of Political Science, Vanderbilt University, 329 Commons Center, Nashville, Tennessee 37203 (bradley.carl.smith@gmail.com, http://bradleycarlsmith.com).

‡Assistant Professor, Department of Political Science, University of Pittsburgh, Posvar Hall, Pittsburgh, PA 15213 (williamspaniel@gmail.com, http://williamspaniel.com).

1 Introduction

Nuclear weapons are a critical independent variable in international relations. Recent research, however, now indicates that the ability to develop nuclear weapons is important as well. Such proficiency determines the domain of potential nuclear states (Sagan (1996, 57-61); Hymans (2006)), alters the speed and cost of proliferation (Sagan (2010); Benson and Wen (2011); Bas and Coe (2016)), incentivizes preventive war (Debs and Monteiro, 2014), and changes nuclear bargains (Spaniel, 2015; Volpe, 2017). Understanding the theoretical significance of latent capacity, scholars have measured this hidden trait using observable characteristics of states. These results have proven fruitful in finding determinants of proliferation, explaining deterrence, and understanding dual-use technology (Fuhrmann (2012*b*); Horowitz (2013); Mattiacci and Jones (2016); Mehta and Whitlark (2016)).

Nevertheless, at present, nuclear proficiency is poorly understood (Sagan, 2010). Aggregation is an inherent challenge. For example, Jo and Gartzke (2007) develop a simple 0 to 7 composite measure of their factors; this cannot distinguish the relative importance of each component in the proliferation process. Meanwhile, Fuhrmann and Tkach (2015) drop the composite measure for a binary variable indicating whether a state has an active uranium enrichment or plutonium reprocessing facility on its soil. Access to fuel is a key hurdle to proliferation (Sagan, 2010) but certainly not the only one. An ideal measure of nuclear proficiency would therefore weight each nuclear-related activity appropriately. Otherwise, the measure may place simple processes on the same level as the most difficult barriers to proliferation.

Reticence is a second challenge. States may forgo nuclear-related activities despite technical capability. Japan, for example, has one of the largest nuclear infrastructures in the world. Few doubt it could develop a nuclear weapon. Yet Japan has not proliferated, while North Korea—a poor and technologically challenged state—tested its first bomb in 2006. Some activities may therefore be noisier indicators of nuclear proficiency than others, independent of their technical difficulty. In turn, even if a measure contained a great number of nuclear activities, the accuracy of proficiency estimates would vary depending upon which specific activities they took part in.

Observability is a third challenge. Nuclear proficiency is an unobservable, conceptual idea. As such, observable consequences can only estimate the underlying characteristic. Using estimated quantities to draw statistical inferences without accounting for this uncertainty therefore risks drawing false conclusions. Indeed, when measurement of an unobservable concept is the goal, a simple indexing scheme prevents the analyst from separating variation due to measurement error from variation in the latent concept (Jackman, 2008). This can have serious consequences for statistical inference, as Treier and Jackman (2008) illustrate in the context of measuring democracy.

Many concepts in the social sciences face similar problems. Scholars have therefore developed tools to solve these observability issues. In particular, studies of legislator ideology (Poole and Rosenthal, 1991), the preferences of Supreme Court justices (Martin and Quinn, 2002), state preferences in the United Nations General Assembly (Bailey, Strezhnev and Voeten, 2013), democratic institutions (Pemstein, Meserve and Melton, 2010), human rights practices (Schnakenberg and Fariss, 2014), and alliance commitments (Benson and Clinton, 2014) adopt a latent-variable approach. These studies treat the inherently unobservable factors as latent traits and learn about them using appropriate statistical models of observed actions. In sum, previous work moves away from the construction of proxy variables in favor of more theoretically grounded measures that allow the data to speak for themselves.

Our paper furthers the study of nuclear capacity through this approach. We create a new dataset of estimated proficiencies called ν-CLEAR. By adopting a latent-variable model grounded in item-response theory (IRT) and taking a Bayesian approach to estimation, ν-CLEAR resolves the aforementioned problems. This statistical model estimates the nuclear proficiencies of states, weighting each activity appropriately in the process. Because our estimation utilizes Bayesian techniques, the model's posterior density provides an uncertainty measurement for each quantity of interest. We also create tools for users to add or subtract component activities from these scores, allowing them to evolve in parallel with new data collection efforts and theoretical advances.

The results reveal new insights on the barriers to nuclear proliferation. By simultaneously estimating proficiency and the relative importance of each activity, our item-response approach contributes to topics of debate among nuclear scholars. Indeed, the estimation process indicates that weapons exploration, reprocessing, and power plants are relatively noisy indicators of nuclear capacity. Put differently, a state *not* engaging in these activities tells us little about its underlying nuclear proficiency. In contrast, the production of nitric or sulfuric acid, although indicative of a lower level of proficiency, provides a much less noisy indicator. All told, these results paint a more nuanced picture of both nuclear proficiency and its determinants.

2 Measuring Nuclear Proficiency

Attempts to measure nuclear proficiency have played a central role in the quantitative literature on nuclear proliferation over the past three decades. However, this literature often lacks clarity (Sagan, 2011), conflating distinct concepts such as nuclear self-sufficiency and capacity. We must therefore clarify the conceptualization of nuclear proficiency that underlies our measurement approach.

To begin, *we define nuclear proficiency as how easily a state may develop nuclear technologies (e.g., power plants or warheads), given their current level of skill and investment of effort.* Thus, a state's "nuclear proficiency" is a composite of skill and effort under our conception. Skill entails a state's ability in nuclear production, independent of the effort invested in obtaining this proficiency. Nuclear skill may arise due to institutional features within a state that facilitate the construction of nuclear technologies or because a state has natural resource endowments that facilitate nuclear development. Effort, on the other hand, entails how hard a state has worked to obtain proficiency. Investment of effort might arise due to a variety of factors, including the availability of nuclear assistance (Kroenig, 2010; Fuhrmann, 2012*a*), the particular preferences of national leaders (Abraham, 2006), and security concerns (Sagan, 1996).

To illustrate, a state intending to build a nuclear deterrent is more likely to construct a uranium enrichment facility. But that state is also more likely to develop enrichment technology for greater levels of underlying skill. Thus, our concept allows proficiency to be increasing in both a state's natural skill in the nuclear arena and its investment into acquiring this proficiency. This understanding of proficiency acknowledges that supply and demand for nuclear weapons technologies are interdependent, influencing a state's unobserved proficiency in an interactive manner. However, regardless of what combination of skill and effort a given state possesses, the result is similar: states with a greater combination of these two factors should exhibit more observable indicators of nuclear proficiency. Measuring this unobservable proficiency by applying a theoretically consistent measurement model is our goal.

As is standard, measurement involves a tradeoff between nuance and feasibility. Researchers ought to know both the strengths and limitations imposed by this conceptualization of nuclear proficiency. For example, our measure is not well-suited for research questions where it is necessary to measure either skill or effort in isolation; if theory points to a relationship between a state's effort in the nuclear arena and some outcome, but *not* skill, our measure is not appropriate.[1] This easy to see by noting that effort to increase proficiency to handle security threats is a primary motivation for an activity like a nuclear weapons pro-

[1] Although a two-dimensional measure is feasible to estimate, we choose to estimate a single proficiency dimension for two reasons. First, a two-dimensional score requires the researcher to assign some variables to one dimension and others to a second dimension. For an example of this in the context of military alliances, see Benson and Clinton (2014). Accordingly, this approach would generate multiple scores for each country-year (one per dimension). As such, it stands in contrast to the existing approaches in the literature, which develop additive indices to assign each country-year a single score. Second, even if we were to separate activities in this way, each dimension would still pick up on a combination of skill and intent. While military and dual-use technologies may be driven by different motivations, intent still drives a state's decision to pursue a given technology. As such, we pursue a one-dimensional approach in this project.

gram. In contrast, active effort does not determine an activity like electricity production, which states develop naturally as they industrialize. If the distinction between skill and effort is not consequential to the study of nuclear weapons, our measure is appropriate. Regardless, this shortcoming is one present in the additive indexing approach as well. Disentangling skill and effort from observed activity is a desirable path for future study, but is outside the scope of the present analysis.

Ultimately, scholars should make the decision of which measurement to use in the context of each specific research question. Our aim is to make this decision easy by detailing precisely how our measure is produced in the following sections.

2.1 Existing Measurements

With this concept laid out and potential concerns addressed, we now describe existing measurement attempts. As seen in Jo and Gartzke (2007) and Fuhrmann and Tkach (2015), the additive-indexing approach unites the literature's most common measures.[2] Using this strategy, scholars collect information on a number of activities related to nuclear weapons production. They then generate a set of dummy variables according to a coding rule for each country-year indicating which activities are observed. The index builds a measure of nuclear proficiency by summing these dummy indicators within each country-year. Thus, the score for a given country-year is simply a count of the number of activities observed.

Assigning weights to each activity addresses the problem of equally valuing each variable. However, it also trades one problem for another: how does one accurately assign weights? Fortunately, the statistical model we implement overcomes this issue by weighting the activities as a part of the estimation process. Our approach therefore "lets the data speak," rather than assigning weights arbitrarily, or by appealing to theoretical arguments about the relative technical importance of various components of nuclear production technology.

A second issue with the additive-indexing approach is that it only produces a single score for each state. To phrase this problem in the language of estimation, this means that this approach is only capable of producing a "point estimate" of the true quantity of interest. Thus, existing measures cannot explicitly account for the inherent uncertainty involved in operationalizing nuclear proficiency (Jackman, 2008). By estimating a statistical model of nuclear proficiency, we again "let the data speak," obtaining a measure of the uncertainty in our estimates. These measures of uncertainty allow for an assessment of the robustness of existing findings that could not account for the unavoidable "noise" in an unobservable trait's measurement.

[2]The Jo and Gartzke dataset extends previous work from Meyer (1986) and Stoll (1996).

2.2 The Item-Response Approach

To overcome the above problems, we adopt tools developed in the psychology literature to assess latent traits (Rasch, 1960). The model belongs to a framework known as *item-response theory*. It uses the responses of individuals to "items"—which may be test questions, roll-call votes, or activities related to the production of nuclear technology—to recover estimates of unobservable quantities of interest. Just as a student's answers to test questions reveal information about their intelligence, or a legislator's voting record reveals information about their ideological preferences, a state's performance on activities related to nuclear production provides information about their underlying nuclear proficiency.

The basic theoretical insight is simple. The higher a state's nuclear proficiency, the more likely it is to engage in activities related to nuclear production. Activities directly related to nuclear production provide clear information. For example, possession of uranium enrichment facilities is clearly indicative of high nuclear proficiency. Indirectly related activities also provide information. For instance, following Jo and Gartzke (2007), we include an indicator of metallurgical capability in our estimation. While metallurgical processing has many uses outside the nuclear arena, it is certainly one component of a state's ability to develop nuclear technology. Thus, we can learn something about a state's nuclear proficiency by observing activities indirectly related to this proficiency as well. Clearly, uranium enrichment and metallurgical processing paint different pictures of nuclear proficiency. The estimates of the item-response model tell us just how much we can learn from each activity.

More specifically, the model uses the data to weight each activity appropriately. It inputs binary response data and smooths the resulting measure of proficiency by automatically accounting for the characteristics of each input. The resulting measure allows for more subtle variation in proficiency.

We now turn to a more systematic description of the model and its parameters. The model we adopt is a straightforward three-parameter item-response model, with a logit link function. While this exercise is different from performing a logit regression, the parameters we estimate roughly correspond to coefficients and intercept terms in a typical logit model. These three parameters allow for the relationship between proficiency and success on each activity to vary in a flexible manner.

Let $i \in \{1, ..., N\}$ index individual states, $t \in \{1, ..., T\}$ index time period, and $k \in \{1, ..., K\}$ index a set of K activities related to the production of nuclear technology. Thus, our data consists of a set of $I \times T \times K$ values, each a 0 or 1, with each particular value indicated by y_{it}^k. A value of 1 indicates participation in activity k by state i in period t, with a value of 0 indicating otherwise.[3] The most important quantity of interest in the model is a state i's nuclear proficiency at time t. In the model, we denote this by the parameter ν_{it}. We interpret this parameter in accordance with our concept of nuclear proficiency outlined previously.

We are also interested in modeling how the activities differ in two important ways. They may vary in the level of proficiency required to attain them, which the parameter α_k captures. They may also separate high and low proficiency states differently, which the parameter β_k captures. Using a logit parameterization, the probability of success in activity k for state i at time t, given the parameters is:

$$\Pr[y_{it}^k = 1] = \text{logit}^{-1}[\beta_k(\nu_{it} - \alpha_k)].$$

This model has the substantive features we desire. Holding ν_{it} fixed, increasing α_k decreases the probability that a state achieves participation in a given activity y_{it}^k. Holding α_k fixed, increasing the proficiency ν_{it} increases the probability of state participation in activity y_{it}^k. More precisely, when $\nu_{it} < \alpha_k$, the probability of success in activity k is strictly below 0.5. As ν_{it} approaches α_k from the left, the probability of success approaches 0.5, with success occurring with probability strictly greater than 0.5 when the state overcomes the "proficiency threshold."

The relationship between β_k and success is more nuanced. β_k captures how cleanly each activity separates high proficiency from low proficiency states. This variation is key, as the activities we include are plausibly related to nuclear proficiency to varying degrees. A simple analogy clarifies this point. Calculus questions distinguish students with high and low math proficiency. However, adding a calculus question to a literature exam would not separate students with high reading proficiency from those with low reading proficiency.

Thus, different activities separate individuals along the latent trait of interest to varying degrees. The parameter β_k accounts for this. Incorporating this into the model is especially important for the estimation of nuclear proficiencies, as we wish to evaluate activities taken as indicators of proficiency as well as proficiency itself.

As β_k increases, the difference between α_k and proficiency becomes more pronounced, influencing the probability of activity participation positively for states that have exceeded the proficiency threshold and negatively for those below. Figure 2 illustrates this effect with predicted probabilities for an activity with $\alpha_k = 0$ across a range of abilities.[4] Each line

[3] Another approach would be to model continuous responses. However, much of the data on activities of interest is in the form of binary responses, such as that of Jo and Gartzke (2007) and Fuhrmann (2012*b*). The main limitation in our view is that use of a binary response approach may produce estimates that are less fine-grained than those resulting from use of a more continuous response.

[4] Note that $\alpha = 0$ is an arbitrary choice here and does not influence the substantive point of the graph.

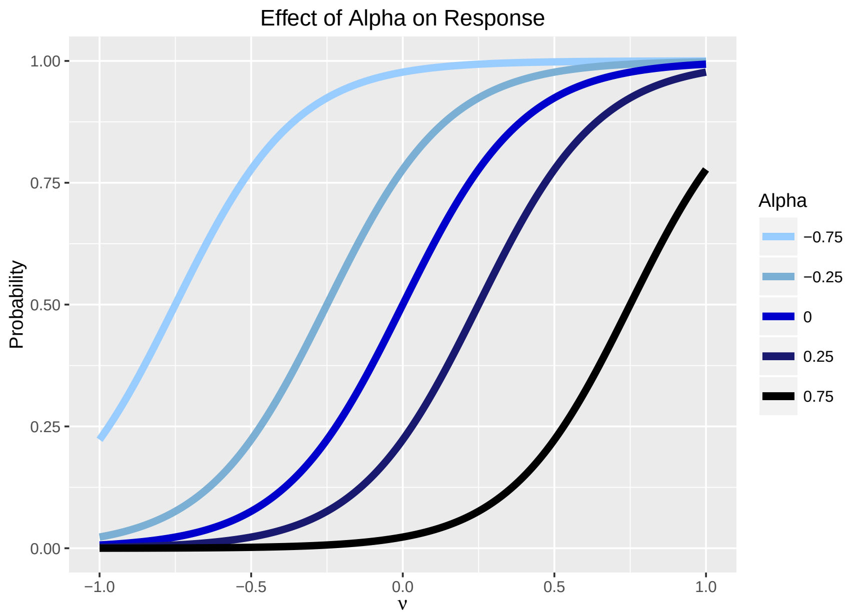

Figure 1: The influence of variations in α on the relationship between proficiency and participation in an activity with $\beta = 5$.

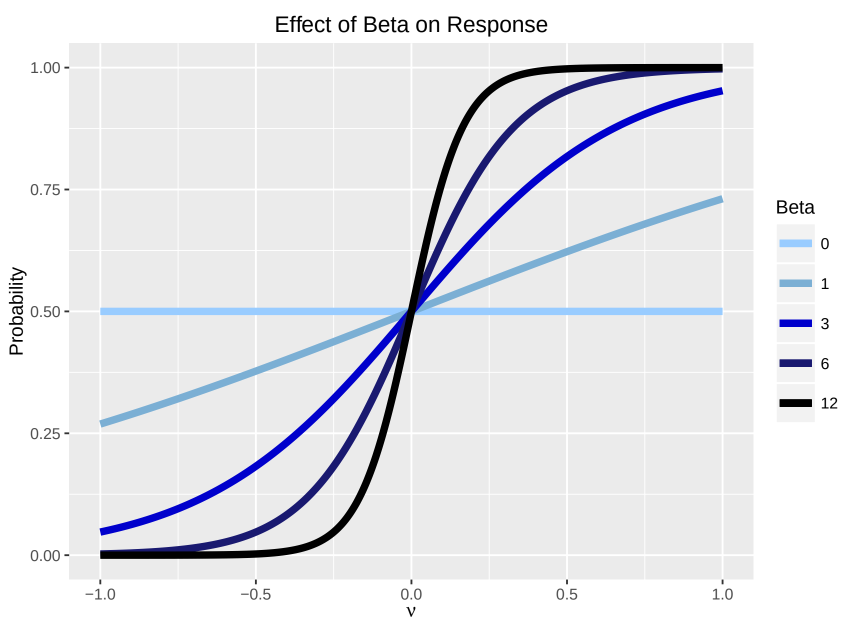

Figure 2: The influence of variations in β on the relationship between proficiency and participation in an activity with $\alpha = 0$.

represents a different value of β. Note that as β increases, the rate at which a state's proficiency influences the probability of participation increases. When $\beta = 0$, all states behave the same, regardless of proficiency. In contrast, for $\beta = 12$, high ability states participate in an activity almost deterministically. Similarly, for high β, the lowest ability states almost never participate. This means that for high values of β, factors other than a state's proficiency have little effect on the probability of participation.

As such, β parameterizes how deterministic the relationship between proficiency and activity participation is. A value of β close to 0 indicates that the relationship between proficiency and participation is very noisy. In contrast, as β increases, the relationship becomes sharper, with the probability that relatively high and low states participate in a given activity approaching 1 and 0, respectively. Some activities are noisier indicators of nuclear proficiency than others. As such, our estimates for β contribute to weighting each activity appropriately by accounting for this variation.[5]

Before covering our data and estimation, we pause to describe two features of the approach. First, we highlight the model's key assumption: *local independence*. Independence is important here, as it is well established that the steps to proliferation are related to one another. Local independence requires that the probability of success on any two activities are independent, *conditional on the parameters*. This is a potential concern given the relationship between activities included in the measure. For example, participating in one activity may make another less attractive, which would be indicated by negative correlation in the raw data. We find no evidence of this, as discussed in the following section.[6] However, these potential violations are worth note. Development of a new statistical approach to account for this is one avenue for future research that we return to in the conclusion.

Second, it is important to note that this model does not "stack the deck" in favor of rejecting the assumptions of the additive-indexing approach. In fact, if the implicit assumptions of the additive-indexing approach hold in the data generating process, then the estimated α_k and β_k terms will produce estimates with a relative ordering of each ν_{it} that is identical to

Choosing another value, such as $\alpha = 1$, would simply shift the point at which each predicted probability crosses the threshold of 0.5 to the right.

[5]For our analysis, we constrain each β_k to positive values. This normalization is analogous to setting a "direction" in models of roll-call voting, assigning a left-right ordering to voters. As such, this does not influence the substantive results of the estimation. We have also estimated the model without this constraint and the results remain unchanged.

[6]For example, possession of uranium enrichment facilities may reduce the desirability of reprocessing plants. As such, one may expect a negative correlation between enrichment and reprocessing in the data, indicating a violation of this assumption. However, we find that there is substantial overlap in these two activities, and that they are positively correlated. In particular, using data from Fuhrmann and Tkach (2015), we find that nearly 75% of states with an enrichment facility also have reprocessing capability. Similarly, 61% of states with reprocessing capabilities also have an enrichment plant.

the ordering of existing additive indices.

3 Data & Estimation

For our initial estimation of capacity scores, we incorporate data from a number of sources. By including a wide range of variables, we allow the model to draw out the relevant information through estimation. This is a strength of our approach. Our model frees us to include a large set of variables plausibly related to nuclear proficiency and weights them appropriately in the resulting estimate.

Because we desire proficiency estimates that can address a broad range of substantive questions, we estimate two versions of proficiency scores. Here, we distinguish between "infrastructure" indicators and "weapons" indicators. Using only infrastructure indicators (e.g., uranium possession and plutonium reprocessing) generates a measurement closest to what the literature describes as nuclear latency. The second measure also includes indicators for weapons programs, nuclear weapons tests, and possession of nuclear weapons.

Given prior research on nuclear latency, one may wonder why we develop the second measure. This literature emphasizes that latency measures are useful because they provide information of "how quickly a state could develop a nuclear weapon from its current state of technological development if it chose to do so" (Sagan, 2010, 90). Our infrastructure measurement generates a theoretically-informed version of this variable.

However, nuclear proficiency persists after acquisition of nuclear weapons. Many research questions therefore would benefit from a more inclusive measure that uses nuclear weapons activities to estimate proficiency. For example, if a researcher wanted to disaggregate the effect of actually possessing a nuclear weapon from the effect of proficiency, he or she would want to include the full measure in a regression.[7] Also as an independent variable, the full measure is useful for quantifying the cost of continuing a program, estimating the effectiveness and reliability of a program. As a dependent variable, researchers can use it to investigate whether there is a "sweet spot" of nuclear proficiency (Volpe, 2017) and whether norms against nuclear technology in general lead to a reduction in mastery.

We list each of the indicators below, according to their descriptions in their original dataset (where applicable). The remaining analysis uses the inclusive measure unless otherwise stated.

Infrastructure Variables

[7]That is, in a regression with just nuclear weapons possession as an independent variable, the corresponding coefficient captures both the effect of having a nuclear weapon and the effect of being proficient enough to build a nuclear weapon. Including both ν and weapons possession can separate the two effects.

- **Uranium Possession:** Indicator of uranium possession. It takes a value of 1 if a country is known to have uranium deposits or has produced uranium in previous years. This and the following six variables come from Jo and Gartzke (2007). All of Jo and Gartzke's variables are coded as 1 in all years after a state first hits the mark.[8]

- **Metallurgical Capability:** Indicator of capability to produce crude steel or aluminum. This provides evidence of state ability to process uranium ore.

- **Chemical Capability:** Indicator of capability to produce nuclear munitions, specifically nitric or sulfuric acid.

- **Nitric Production:** An additional indicator of nuclear munitions capability, measuring whether a country produces non-organic fertilizer.

- **Electricity Production:** Indicator of capability to produce electricity sufficient to run nuclear weapons programs, specifically the capability to generate at least 200 megawatts.

- **Nuclear Engineering:** Indicator of whether a state has at least three reactor-years of of experience.

- **Explosive Production:** Indicator of capability to produce electronics necessary for explosive detonation, specifically whether a country manufactures motor vehicles or assembles motors and produces televisions or radios.

- **Heavy Water Reactor (HWR) Power Plant:** Indicator of whether a country had a civilian power plant in commercial operation for the given year that used heavy water in its fuel cycle.[9] 14 countries have constructed such plants. We coded all data on civilian power plants from the IAEA's reference series.[10]

- **Non-HWR Power Plant:** Indicator of whether a country had a civilian power plant in commercial operation for the given year that did not use heavy water in its fuel cycle.[11] 34 countries have constructed non-heavy water plants.[12] 11 of these also had a heavy water plant at some time.

[8] For more detailed information with respect to the coding of each of these indicators, see http://pages.ucsd.edu/~egartzke/data/jo_gartzke_0207_codebk_0906.pdf

[9] A power plant's contribution ends upon permanent shutdown.

[10] http://www-pub.iaea.org/MTCD/Publications/PDF/RDS2-32_web.pdf

[11] Such plants can instead use graphite, light water, or breeder reactors.

[12] This count includes Yugoslavia and Slovenia as distinct. As such, Serbia (Yugoslavia's successor state as coded by Correlates of War) lost its power plant in 1992 after Slovenia declared independence.

- **Uranium Enrichment:** Indicator of whether a state operated a plant in the given year that enriches uranium.[13] This category covers a wide-variety of enrichment strategies, including diffusion and centrifuge techniques. Data on enrichment and reprocessing below come from Fuhrmann and Tkach (2015).

- **Reprocessing:** Indicator of whether a state operated a plant in the given year that reprocesses uranium previously used as fuel.[14]

- **Submarines:** Indicator of whether a country had a commissioned nuclear-powered submarine (constructed by that country) in the given year. Five countries meet this criterion: the United States (USS Nautilus, 1954), the Soviet Union (Leninsky Komsomol, 1958), the United Kingdom (HMS Dreadnought, 1963), France (Redoutable, 1971), and China (Type 091, 1974).[15]

Weapons Variables

- **Weapons Exploration:** Indicator of whether a state is exploring nuclear weapons. Drawn from Bleek (2010)

- **Weapons Pursuit:** Indicator of whether a state is actively pursuing nuclear weapons. Also drawn from Bleek (2010).

- **Nuclear Test:** Indicator of whether a country has a confirmed successful test of a nuclear bomb that year or in any year prior.

- **Nuclear Weapons:** Indicator of whether a country is a *de facto* nuclear weapons state. While many studies use latent measures as a means of predicting proliferation, we include possession of nuclear weapons to allow the model to show how much more capable these states are. Note that there is variation between possession and testing; India went more than two decades between detonating the "Smiling Buddha" and producing a bomb, while Israel and South Africa never had a confirmed test.[16] Data come from Jo and Gartzke (2007).

[13] Natural uranium contains only 0.72% of the U-235 isotope; enrichment concentrates the U-235. Low-enriched uranium (with concentrations of U-235 under 20%) can fuel light water reactor power production; highly enriched uranium (with concentrations of U-235 exceeding 90%) are necessary for uranium-based weapons.

[14] Reprocessing collects plutonium, which states can use for a plutonium-based nuclear weapon or as a fuel for certain types of reactors. It also reduces the quantity of nuclear waste created through the fuel process.

[15] This excludes India, which has leased a Soviet and Russian submarine. Both India and Brazil have native submarine programs in development.

[16] There is speculation that the "Vela Incident" was the result of a nuclear detonation cosponsored by the two countries.

Figure 3 illustrates the correlation between each pair of these variables. Importantly, this correlation table indicates that each of the activities are positively correlated, although some only weakly so. This is reassuring for our approach. In particular, this positive correlation is consistent with each activity being partially driven by a common, underlying trait.

Although these indicators provide a reasonable start, we do not claim that they exhaust every possible indicator of nuclear proficiency. Scholars will undoubtedly code new nuclear proficiency factors, and others may wish to add other variables for specific research questions (e.g., covert nuclear assistance). We therefore provide a simple estimation procedure that can incorporate additional data. Our user-friendly code allows scholars to utilize this approach to develop alternative formulations of nuclear proficiency scores. All that is necessary to modify the scores we present is to provide a properly formatted dataset of activity indicators.

By providing this code, we hope to contribute to the debate on the determinants of nuclear proficiency by providing both transparency and clarity. Data for many important nuclear factors are secretive in nature. Thus, if users foresee problems with the scores we provide and have newly available raw data, this code will allow for straightforward reestimation in subsequent studies.

To obtain a proficiency score for each country-year, we require the estimation of a large number of parameters. With 8086 country years and 15 activities, we have 8118 parameters to estimate. Consequently, we adopt a Bayesian approach that is well-suited to problems of this kind. We estimate the model using the Stan platform, as implemented in the RStan package (Carpenter, Gelman and Hoffman, N.d.). Stan performs estimation via Hamiltonian Monte Carlo. The results presented in the following sections come from three parallel chains, each with 10,000 draws after a burn-in period of 5000 draws. Standard diagnostics indicated convergence.

The scale, location, and direction of the parameters are not well-identified without some normalization.[17] As such, the analyst must constrain the model to produce well-behaved estimates. To fix these various identification issues, we imbue the parameters with the following priors:

- $\alpha \sim N(0, 3)$
- $\beta \sim \text{lognormal}(0, 1)$
- $\nu \sim N(0, 1)$

[17] Without this, the model suffers from rotational invariance. This problem arises without an arbitrary direction specified, as multiplying all of the parameters by -1 does not cause the likelihood to change. In the context of roll-call voting, this issue arises because the direction of liberal and conservative ideologies is arbitrary. Similarly, the posterior distribution can be "scaled" up or down by spreading out the estimates to maintain the mean but increase the standard deviation with no effect on fit.

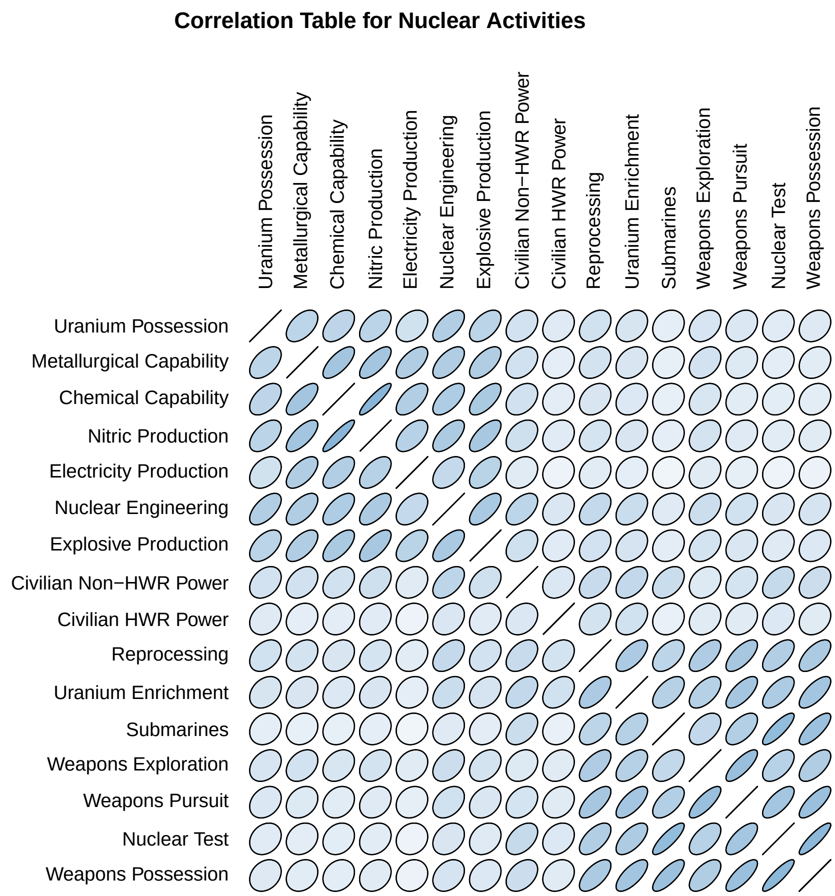

Figure 3: The correlations of each component in our model. Sharper, darker ovals indicate correlations closer to 1, while lighter circles indicate correlations closer to 0. No variables are negatively correlated.

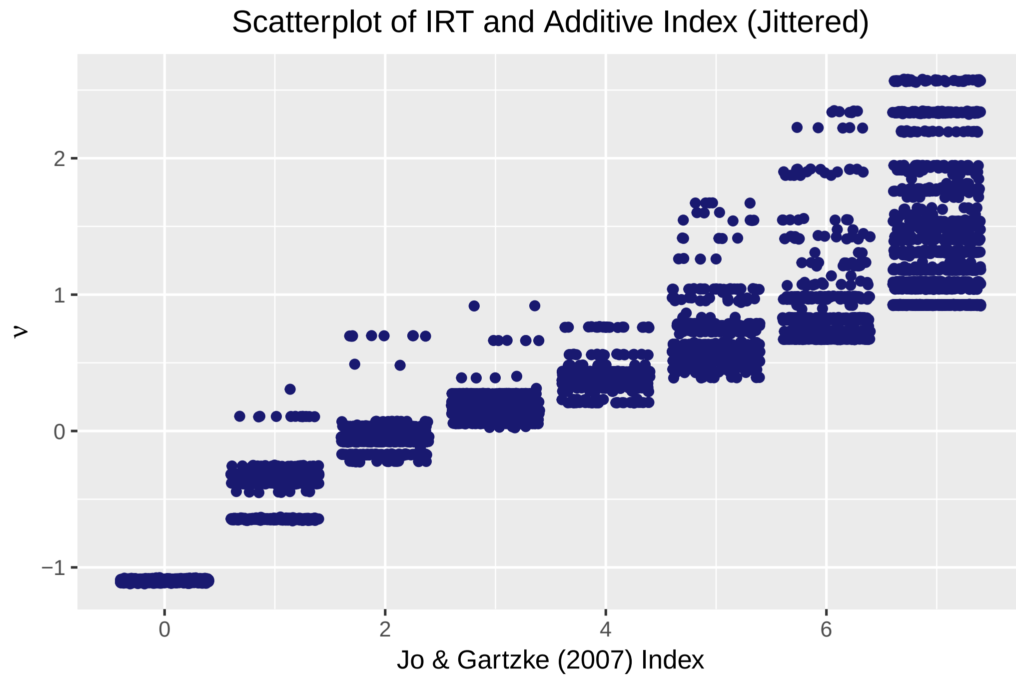

Figure 4: Comparison of ν and Additive Index.

These priors identify the scale, location, and direction of the parameters, allowing for estimation. The normalizations also aid in interpretation of the results. For example, the prior belief on ν centers the posterior with a mean at 0 and a standard deviation of 1. Thus, if the point estimate for a state's proficiency is near 0, that state is of roughly average proficiency.[18] However, this prior does not constrain the estimates to have a normal distribution. In our presentation of the results, we demonstrate that the posterior distribution is consistent with substantive knowledge of the history of proliferation.[19]

4 State-Level Results

Figure 4 presents a first-pass illustration of the relationship between the estimates of ν and an additive index. To construct this additive index, we adopt the approach taken in Jo and

[18]Here, we use the mean of the posterior to calculate point estimates for each parameter.

[19]While the lognormal prior constrains each β to be positive, this is not strictly necessary due to the nature of the data that we implement. Reestimation of the model without this restriction returns substantively identical results, producing a posterior for which each β distribution is well-above 0. Thus, the problem of reflection does not arise here as it may in other political science applications, such as roll-call votes, in which on any given vote, increasing the ideology estimate will result in an increased probability of a yes vote for some legislators and a decrease for others.

Gartzke (2007), summing across their indicators. Note that the x-axis points in this plot are jittered horizontally to aid visualization of the discrete additive index. This plot provides evidence for the face-validity of our measure. While there is variation from the additive index, our estimates share broad features. No points inhabit the upper-left or lower-right of the graph, indicating that our scores positively correlate with the additive index.

While the similar predictions of these two approaches indicates that our item-response estimates have face validity, the differences between the two scores are also informative. Consider how, within each non-zero category along the along x-axis of Figure 4, there is substantial variation along the y-axis. This demonstrates that our model captures variation within each category missing from Jo and Gartzke's measure. Note the substantial overlap in the estimated values of ν among observations assigned index values 5 through 7. Many observations with an index value 6 receive lower ν values than observations with an index value of 5, while others received ν values higher than observations assigned an index value of 7.[20]

This overlap is consistent with variation in the appropriate weighting of each activity. By weighting each activity equally, the additive index approach implicitly assumes that two states given a score of 6 are equally capable, even if those scores were generated from different combinations of activities. However, our procedure can separate these scores, assigning different values to different activities in terms of their influence on estimation of the trait.[21]

Due to concentrations of particular levels of proficiency, the scatterplot obscures the precise distribution of our scores. We therefore include a histogram of ν in Figure 5. Note that the modal state is at the lowest end of our scale, indicating a concentration of states with low proficiencies. In contrast, the most capable states on the right end of the scale are infrequent. States in the middle range are somewhat common and are distributed throughout the interval.

Both the scatterplot and histogram give no historical context or indication of how proficiency has changed over time. They simply record one observation per country-year. While those looking for full state-level variation across time can consult our dataset, Figure 6 provides a snapshot of the year 2001. When viewed in the context the histogram of Figure 5, which pools all country-years, worldwide ν scores are generally high at this point. Europe uniformly has strong capacity, while most of Asia and South America rank high compared to historical averages. Only Africa lags behind.

This picture of high worldwide levels of nuclear proficiency is consistent with ongoing

[20]To ensure that these differences were not driven by new data sources, we estimated ν with only the components of Jo and Gartzke's (2007) original measure. The scatterplot produced by this approach in the appendix demonstrates even more deviation from the additive index.

[21]As a result, our method points to variation that was lacking in the original measure. For example, the UK and Colombia are assigned the same score under Jo and Gartzke's (2007) index in the mid-1970s. However, our measure ranks the UK considerably higher, assigning a score that is distinguishable from Colombia's even when uncertainty in the posterior distribution is accounted for.

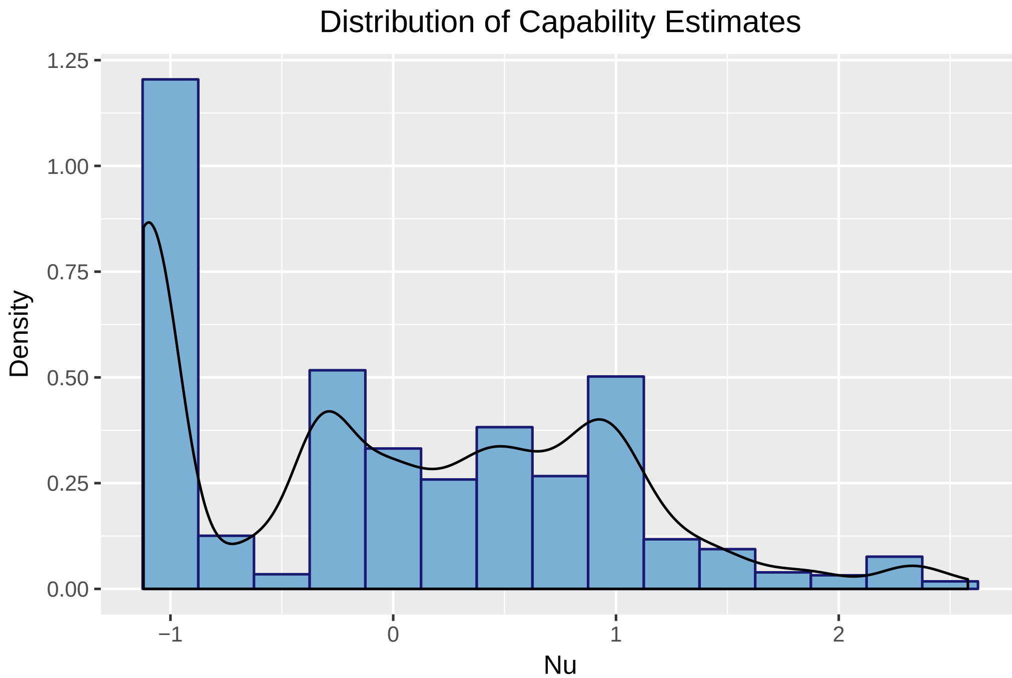

Figure 5: Density and histogram of ν-CLEAR estimates across the entire temporal range.

concern that the spread of nuclear technology will inexorably lead to a large number of nuclear weapons states. Our proficiency estimates provide insight here. By comparing our proficiency scores over time to a benchmark proliferating state and its score, we can consider whether the trend of increasing proficiency is likely to correspond to a trend of nuclear weapons proliferation.

To establish a benchmark state, a particularly astute observer of Figure 6 might notice North Korea's middling ν score for 2001. North Korea performs poorly even though Kim Jung Il's regime was just five years away from testing its first nuclear weapon.[22] To some, this may be cause for concern: if North Korea was so close to developing a nuclear weapon at this time despite such a low score, does this mean that many other countries could develop a weapon in short order?

Based on observable characteristics, our ν scores indicate this is so. Nevertheless, such an outcome appears unlikely. Using North Korea's 2001 ν score as a baseline threshold for proliferation proficiency, Figure 7 plots an estimate of states that exceed this threshold over

[22]North Korea is an especially apt state to use for this purpose, as it allows us to establish a lower bound on proficiency necessary for states to produce a weapon in short order. To allay concerns with this approach, we reproduce Figure 7 with South Africa's proficiency in 1977 and 1981. These plots appear in the appendix and demonstrate the same pattern; there are many more high proficiency states than states that possess nuclear weapons.

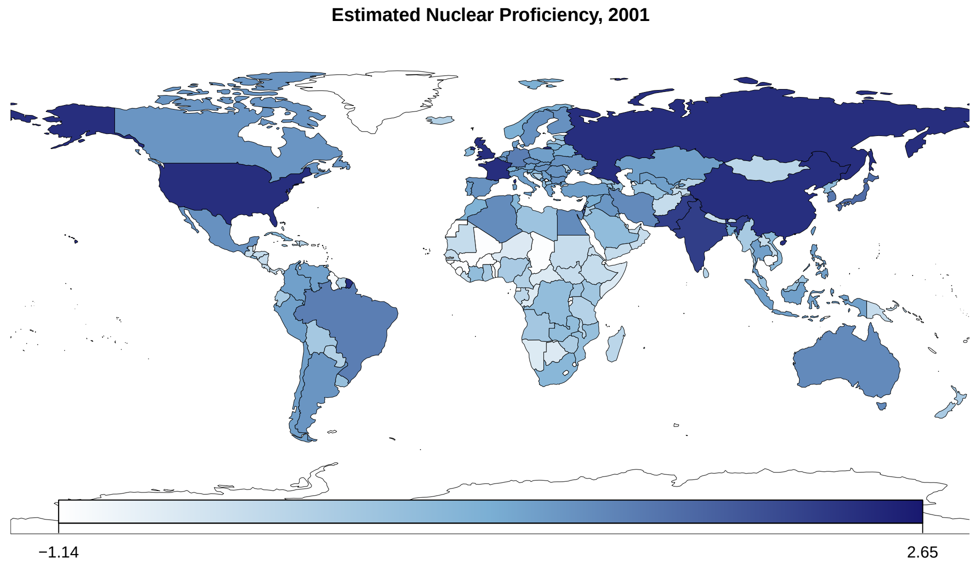

Figure 6: Estimated nuclear proficiencies in 2001 according to ν-CLEAR. Darker shades indicate higher estimates of ν.

time. This group grows rapidly from the end of World War II to the late 1960s and then alternates between no growth and low growth. Thus, a large group of states with relatively high nuclear proficiencies is not a recent phenomenon; capacity has been high for a long period of time.

This is cause for optimism. Figure 7 also tracks the number of states with nuclear weapons over the same period. Despite the explosion of potential proliferators in the 1950s and 1960s, the number of nuclear weapons states increased at a slow rate.[23] Security, institutions, and bargaining theory help explain the discrepancy. States with the proficiency to develop a weapon may elect not to if the start of a program would elicit preventive war (Chadefaux, 2011; Debs and Monteiro, 2014). The Nuclear Nonproliferation Treaty (NPT) has codified international commitments to nuclear reticence, raising the cost of violation and reducing proliferation (Coe and Vaynman, 2015). The NPT also provides concessions to signatories, and rivals can use other inducements to raise the opportunity cost of proliferation (Spaniel, 2015). These successes mean that the discrepancy seen in Figure 7 is not an accident, though the high levels of global nuclear capacity mean that the nonproliferation regime must remain vigilant.

This plot also demonstrates our ability to incorporate uncertainty into the analysis of nuclear proliferation. While the simple count of states with a higher ν score than North Korea

[23]Nuclear weapons scholars are likely familiar with this type of figure from Hymans (2006, 4). Our estimation procedure predicts a similar number of potential proliferators.

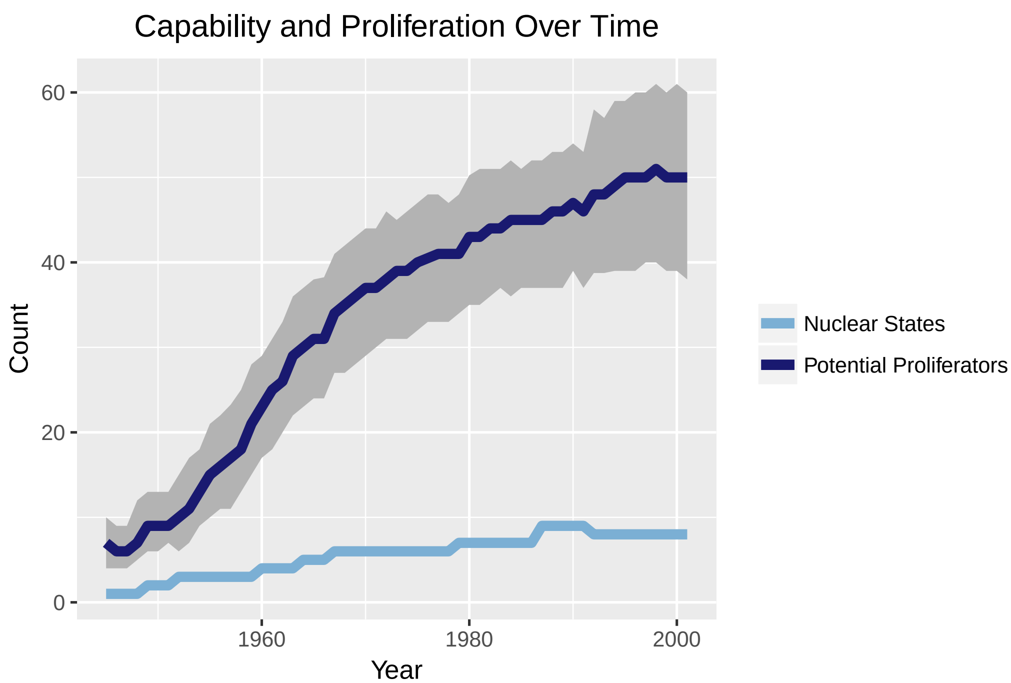

Figure 7: Nuclear weapons states versus potential proliferators across time. The shaded region reflects the uncertainty in the posterior distribution.

in 2001 provides some information, it is worthwhile to consider whether this measurement is robust to the uncertainty in the scores. The confidence interval shown in the plot demonstrates that it is. To create this interval, we simulate draws from the posterior, and in each draw calculate the number of states with a score higher than North Korea in 2001. After repeating this process 1000 times, we take the results of these simulations and for each year, plot the median and a 95% interval around it. Thus, this plot demonstrates that the finding is robust to an explicit correction for this uncertainty. There are also proficiency gaps between proliferators and nonproliferators. In 1982, for example, South Africa ranks below Brazil, India, and Pakistan despite the latter three countries not yet possessing a nuclear weapon. These discrepancies only become more pronounced when looking at the infrastructure measure instead. Without weapons pursuit to bolster North Korea's score, 87 countries rank ahead of it in 2001; without weapons possession to bolster South Africa's score, 67 countries rank ahead of it in 1982.

With macro-trends out of the way, we turn to dyadic comparisons and the evolution of particular scores over time. Estimating proficiencies directly rather than using a simple index clarifies comparisons among states at various stages of the proliferation process. If each indicator in the index is not related to the underlying quantity of interest in the same

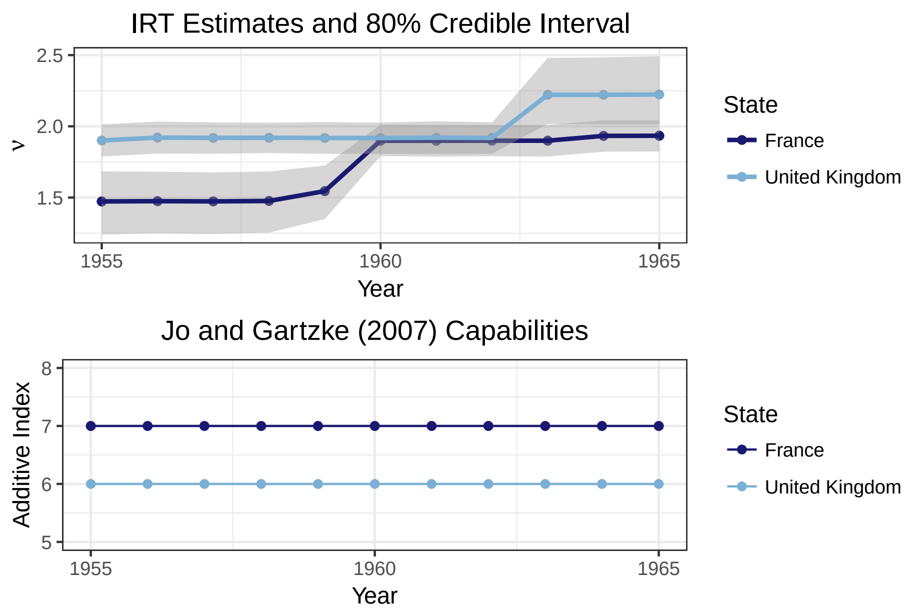

Figure 8: Comparison of France's and the United Kingdom's proficiencies from 1955 to 1965.

way, then a sum that does not weight each indicator uniquely may produce poor estimates. Additionally, a sum of indicators carries no information about the analyst's uncertainty over how well this sum maps onto the trait of interest. These shortcomings of the additive-indexing approach result in a number of roadblocks to inference of the true underlying quantity.

For example, an additive index could assign a clear rank ordering over the traits associated with a pair of states when it should not. Figure 8 compares the nuclear proficiency scores Jo and Gartzke (2007) provide for France and the UK from 1955 through 1965. For the entire time period, France's score under the additive indexing approach is consistently higher than the UK's. Due to this, an analyst might conclude that France's proficiency outranked the UK's during this time. Under ν-CLEAR, beginning in 1960, the rank ordering becomes much less clear. Indeed, the two procedures produce different pictures of nuclear proficiency.[24]

In addition to cross-country comparisons, we can also consider temporal variation for a single state and check whether ν captures important within-country variation. Figure 9 plots Argentina's proficiency estimate across 1955 to 2001. Argentina rises steadily to the mid-1970s, with a short drop occurring due to the closing of reprocessing facilities in 1973. Within a few years, Argentina's proficiency again rises, first by initiating a formal exploration into

[24]A potential concern with this result is that these substantive differences may have arisen because our estimates of ν incorporates indicators not in the original Jo & Gartzke index. However, if we estimate ν using only these 7 indicators, the same pattern emerges. A reproduction of Figure 8 demonstrating this appears in the appendix.

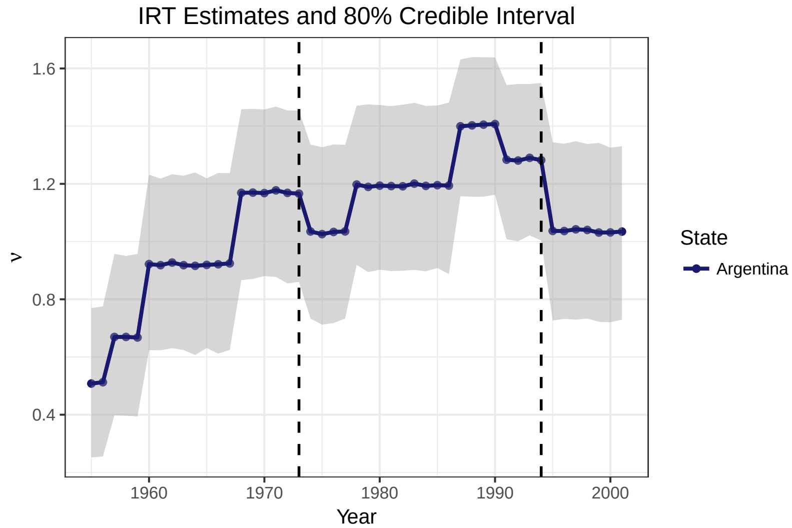

Figure 9: Proficiency estimates for Argentina show temporal variation consistent with its exploration into nuclear weapons. We have highlighted two important points that our measure captures with dotted vertical lines. First, in 1973 a short drop in proficiency corresponds to the closing of reprocessing facilities. A drop also occurs at the second dotted line in 1994 when Argentina ends its foray into enrichment.

nuclear weapons and reaching a peak with the development of an enrichment facility in 1987. With its weapons exploration efforts ceasing after the signing of the Mendoza Declaration in 1991, Argentina's proficiency falls. It decreases again when Argentina ended its foray into enrichment in 1994. This demonstrates that our estimates of proficiency track with state efforts at weapons production. When effort ceases, proficiency correspondingly drops.

5 Activity Characteristics

Another advantage of the IRT framework is that it allows for a more nuanced understanding of how the trait of interest relates to multiple activities. The additive indexing approach relies implicitly assumes that each activity should be weighted the same in terms of contribution to the trait. The α and β parameters in our model relax this assumption. Inspection of these estimates contributes to debate about the ranking of various activities in the proliferation process.

First, consider our estimates of each α_k. This parameter can be interpreted as the "proficiency threshold" of item k. When a state's proficiency, ν, is exactly equal to the associated α_k of an activity, that state will engage in the activity with probability 0.5. When $\nu > \alpha_k$, the state engages with probability above 0.5; likewise, when $\nu < \alpha_k$, the state engages with probability below 0.5.

With this in mind, our model draws interesting results on the relative proficiency indicated by observation of each activity. Table 1 summarizes the results. The coefficients for each α mean are comparisons to 0, where 0 represents the proficiency of the average state for the category in question. Thus, the positive coefficient on uranium possession indicates that the average state is unlikely to fulfill its requirements; in contrast, the negative coefficient on electricity production indicates that the average state is likely to.

Larger values for α indicate activities that require higher levels of proficiency. Consequently—and intuitively—the model shows that nuclear weapons tests, nuclear weapons possession, and heavy water power facilities rank higher on this trait. Having a nuclear program, an enrichment facility, or non-heavy water nuclear power plants falls in the middle tier. In contrast, the seven components of the Jo and Gartzke scores rank as the seven categories with the lowest values of α in our model. This indicates that limiting the discussion to just these components misses out on the critical factors that play a greater role in revealing nuclear capacity.

Table 1: Estimates of α for each activity

	Mean	10th Percentile	90th Percentile
Electricity Production	-0.41	-0.43	-0.39
Metallurgical Capability	0.16	0.14	0.18
Chemical Capability	0.20	0.18	0.22
Nitric Production	0.35	0.33	0.37
Explosive Production	0.43	0.41	0.45
Uranium Possession	0.59	0.57	0.62
Nuclear Engineering	0.62	0.60	0.64
Reprocessing	1.44	1.41	1.48
Exploration	1.46	1.43	1.50
Non-HWR Power	1.53	1.49	1.57
Uranium Enrichment	1.53	1.50	1.57
Pursuit	1.62	1.58	1.66
Nuclear Weapons	1.74	1.70	1.77
Nuclear Test	1.82	1.78	1.86
Submarines	2.01	1.97	2.05
HWR Power	2.44	2.35	2.53

Consider the estimates of each β_k. Recall that β_k parameterizes how deterministic the relationship between proficiency and activity participation is. For higher values of β_k, states that exceed the relevant threshold on proficiency, α_k, become more likely to participate, while states that fail to meet this threshold become less likely to participate. In contrast, for low values of β_k the relationship between proficiency and participation is noisier.

Table 2: Estimates of β for each Activity

	β mean	10th Percentile	90th Percentile
Uranium Possession	2.02	1.95	2.09
HWR Power	2.13	1.99	2.27
Non-HWR Power	2.81	2.67	2.96
Exploration	3.11	2.95	3.28
Explosive Production	3.63	3.50	3.77
Metallurgical Capability	4.25	4.08	4.42
Pursuit	4.31	4.03	4.59
Reprocessing	4.76	4.44	5.09
Nuclear Engineering	4.84	4.63	5.06
Uranium Enrichment	4.97	4.64	5.33
Electricity Production	5.00	4.71	5.28
Nitric Production	5.79	5.51	6.09
Chemical Capability	5.81	5.52	6.11
Nuclear Weapons	14.56	11.99	17.54
Nuclear Test	23.59	16.74	32.00
Submarines	30.50	18.35	45.28

With this, Table 2 presents our estimates of each β_k. Note that chemical proficiency and nitric production, despite being relatively low in terms of α, score high on β. Thus, states exceeding this threshold are particularly likely to engage in these activities. This is consistent with our interpretation of nuclear proficiency, as high β indicates that states that can engage in these activities do so. Both chemical and nitric production are beneficial activities, even outside the realm of nuclear production. Similarly, these activities are not likely to attract unwanted attention from the international community, so states have little incentive to avoid them. Thus, the relationship between proficiency and participation is not very noisy.

In contrast, the operation of enrichment and reprocessing facilities, while having a higher α_k estimates, have middling values of β_k. Even when a state's nuclear proficiency is high, barriers besides nuclear proficiency provide roadblocks to the management of these sites. Enrichment facilities are routinely identified as a key factor for many states' paths to development. Consequently, factors beyond proficiency influence the construction of these facilities. Thus, unobserved nuclear proficiency influences both a state's chemical proficiency and nitric production comparatively more. Put differently, chemical or nitric production provides the analyst with a relatively deterministic indicator of a low level of proficiency. In contrast, although enrichment and reprocessing are associated with relatively high proficiency, the relationship is noisier, as high proficiency states may forgo these activities.

Additionally, the high β_k estimate for nuclear testing provides evidence that our estimates of ν are valid indicators of nuclear proficiency. The high β estimate on this activity means that states exceeding the proficiency threshold for weapons tests are likely to do so. As nuclear weapons tests are clearly related to nuclear proficiency, this provides assurance of the validity of our estimates across all activities. Because an activity clearly related to nuclear proficiency produces intuitive estimates, we are confident that the remainder of our estimates accurately capture nuclear proficiency.

6 Replication

Finally, we demonstrate the utility of our approach by providing a check on the robustness of two existing results that tie nuclear proficiency to proliferation and conflict initiation. As we have argued, nuclear proficiency is a quantity of interest that is inherently unobservable. Thus, measures of nuclear proficiency estimate this quantity. Unless analysts account for the corresponding uncertainty, they risk drawing unwarranted conclusions induced by over-precision. Using ν incorporates this uncertainty into the estimation to guard against this concern.

First, we replicate Model 1 from Jo and Gartzke (2007). This exercise is one example of the usefulness of our infrastructure estimate of ν. As the outcome variable in our replication is the presence of a weapons program, use of the more inclusive measure, which includes program status, would induce bias.[25] While there are a number of conflicts between ν and Jo and Gartzke's additive index, the measures still correlate at roughly .9. However, our results indicate that substantial uncertainty remains for many country-years. To reevaluate Jo and Gartzke's claims in light of this uncertainty, we perform a nonparametric bootstrap using the posterior distribution of our estimates of ν.[26] In the first step, we draw one observation of ν for each state from the posterior distribution. Next, we merge these scores into the replication data from Jo and Gartzke (2007). With this, we draw a sample (with replacement) of these data equal to the number of observations. Finally, we perform a logit regression and store the coefficients. This process is then repeated 10,000 times.

Table 3 presents the results of this estimation, along with a replication of the original result. For the bootstrapped model that appears in the first column, the reported coefficient estimates are the means obtained from the stored bootstrap coefficients, and the standard errors are the respective standard deviations of the stored estimates. Note that the coefficient on ν, our

[25] As a check on our own robustness, we apply the ν-CLEAR procedure to *only* the seven indicators used in Jo and Gartzke's (2007) original index. When this is used, we the original result again replicates. See the appendix.

[26] For a similar use of this approach, see Ashraf and Galor (2013). An alternative method for incorporating this uncertainty is to use a procedure typically utilized for multiply imputed missing data, as in Schnakenberg and Fariss (2014). This procedure uses Rubin's (2004) formula to calculate standard errors and point estimates.

Table 3: Replication of Jo and Gartzke (2007) Using ν-CLEAR

	Dependent variable:	
	Weapons Program	
	(1)	(2)
ν	1.014*** (0.102)	
Jo & Gartzke Measure		0.484*** (0.079)
Economic Capacity	0.872 (1.265)	1.483 (1.944)
Diffusion	0.851*** (0.178)	1.055*** (0.251)
Conventional Threat	0.697*** (0.108)	0.700*** (0.258)
Nuclear Rival	−0.837*** (0.238)	−0.914*** (0.364)
Nuclear Defender	0.171 (0.141)	−0.098 (0.306)
Diplomatic Isolation	−0.007 (0.222)	−0.060 (0.438)
Domestic Unrest	−0.059 (0.071)	−0.148** (0.096)
Democracy	−0.018 (0.014)	−0.026* (0.022)
NPT Ratification	−0.589*** (0.182)	−0.781*** (0.363)
NPT (system effect)	0.006** (0.003)	0.005* (0.004)
Major Power	1.856*** (0.297)	2.000*** (0.388)
Regional Power	1.530*** (0.183)	1.549*** (0.236)
Count 1	−0.118*** (0.017)	−0.113*** (0.012)
Constant	−4.070*** (0.565)	−6.354*** (1.001)
Observations	4,697	4,697
Log Likelihood	−273.505	−256.712
Akaike Inf. Crit.	577.010	543.424
Note:		*p<0.1; **p<0.05; ***p<0.01

alternative indicator of nuclear proficiency, remains positive and statistically significant.[27] The signs and significance for the controls are generally consistent as well.

Next, we consider a second result from Fuhrmann and Tkach (2015). Again, we utilize our infrastructure variable of ν, as it allows us to include a separate indicator of weapons possession, as in the original model. Their analysis investigates how a state's possession of an enrichment or reprocessing facility impacts the likelihood that it is targeted in a militarized interstate dispute.[28] The authors find that opponents are less likely to target high capacity states. In Fuhrmann and Tkach's (2015) original analysis, the independent variable of interest is a binary "latency" variable coded as a 1 if a state possesses enrichment and reprocessing facilities in a given year. However, the authors suggest that future work should assess whether the results hold under a measure that allows for variation in the degree of nuclear proficiency. We perform this exercise using ν.

The replication data provided by the authors is on directed-dyads from 1945 to 2000 and includes a host of standard control variables.[29] Just as in our replication of Jo and Gartzke (2007), we assess the robustness of this finding by replacing Fuhrmann and Tkach's (2015) scores with ν-CLEAR estimates in our probit regressions. Additionally, we employ a non-parametric clustered bootstrap to account for estimation uncertainty.

Looking to the results, we find that ν produces a result that is only partially consistent with Fuhrmann and Tkach's (2015) finding. While the consistent coefficients on both "Latency A" and "ν_A" indicate a positive relationship between a challenger's nuclear proficiency and the initiation of conflict, the picture is different with respect to target states. To see this, compare the coefficients on "Latency B" with "ν_B" in Table 4. In contrast to the finding using Fuhrmann and Tkach's (2015) binary measure, we do not see a statistically significant effect relating the ν score of a target state to conflict initiation. Additionally, the AIC indicates that our model provides an improved fit.

Why do our estimates produce a result different from the original model? One possibility is that the uncertainty in our estimates accounts for uncertainty about the underlying nuclear proficiency, which was not accounted for in the original analysis. Another possibility—and one that Fuhrmann and Tkach anticipate—is that opponents prefer launching preventive war against high capacity states (Fuhrmann and Kreps, 2010). Consequently, even if capacity provides benefits for some (Paul, 2000, 59), this counterveiling effect masks it. A third

[27]In an additional robustness exercise, we account for Montgomery and Sagan's (2009) observation that the original coding of program status in Jo and Gartzke (2007) does not remove states from the risk pool after the first year in which they have a program. This may bias the results in favor of states with long-standing programs. As such, we utilize this alternative coding. The results in this case on our proficiency measure are unchanged. We still recover a positive and statistically significant coefficient on ν.

[28]This measure has a weak correlation of approximately 0.34 with ν-CLEAR.

[29]Details of the controls can be found in the original paper.

possibility is a differing conception of the variable of interest; Fuhrmann and Tkach reset a nuclear power's latency variable to 0 whereas the proficiency measure remains whatever the state's ν score was.

Table 4: Replication of Fuhrmann and Tkach (2015) using ν-CLEAR

	Dependent variable:	
	Conflict Initiation	
	(1)	(2)
ν_A	0.107*** (0.020)	
ν_B	0.020 (0.017)	
Latency A		0.151*** (0.040)
Latency B		−0.133*** (0.047)
Nuclear Weapons A	0.150** (0.063)	0.316*** (0.071)
Nuclear Weapons B	−0.035 (0.068)	−0.053 (0.082)
Nuke A * Nuke B	−0.237 (0.153)	−0.210 (0.136)
Democracy A	0.015*** (0.006)	0.020*** (0.006)
Democracy B	0.040*** (0.005)	0.043*** (0.006)
Democracy A * Democracy B	−0.005*** (0.001)	−0.005*** (0.001)
Rivalry A	0.270*** (0.028)	0.286*** (0.032)
Rivalry B	0.160*** (0.028)	0.162*** (0.030)
Dyadic Rivalry	1.101*** (0.042)	1.116*** (0.051)
Contiguity	−0.140*** (0.029)	−0.137*** (0.044)
ln(distance)	−0.048*** (0.017)	−0.050*** (0.026)
Alliance	0.057 (0.037)	0.044 (0.040)
CINC A	−0.185 (0.595)	0.398 (0.724)
CINC B	1.304* (0.687)	1.972*** (0.864)
CINC A * CINC B	3.060 (10.058)	0.064 (16.132)
Peace Years	−0.062*** (0.007)	−0.061*** (0.008)
Spline 1	−0.0002*** (0.0001)	−0.0002*** (0.000)
Spline 2	0.0001*** (0.00003)	0.0001*** (0.000)
Spline 3	0.00000 (0.00001)	0.00001 (0.000)
Constant	−2.249*** (0.071)	−2.314*** (0.080)
Observations	1,051,218	1,051,218
Log Likelihood	−5,976.087	−5,997.659
Akaike Inf. Crit.	11,996.170	12,039.320
Note:		*p<0.1; **p<0.05; ***p<0.01

A final possibility is that Fuhrmann and Tkach's (2015) measure does not incorporate the wide range of activities of $\nu-$CLEAR. In particular, the measure originally implemented focuses on the importance of enrichment facilities, with the authors arguing that enrichment and reprocessing is a key part of state latency. In contrast, our estimates of α and β indicate that enrichment facilities, while associated with relatively high proficiency, remain a noisy indicator. Thus, our result could differ due to an improved accounting for the complicated and multi-faceted nature of nuclear proficiency. This is consistent with Jackman's (2008) argument that when a single indicator of a latent concept is used, we cannot separate variation due to measurement error from variation in the latent concept.

We do not believe that this result is the "last word" on the relationship between nuclear proficiency and armed conflict. Rather, we have included this replication exercise as a means to illustrate how existing findings may be sensitive to different measurement approaches.

7 Conclusion

Nuclear capacity is a central concern to both scholars and policymakers. However, existing measures suffer from a number of problems due to the inherently unobservable nature of nuclear proficiency. This paper has adopted an item-response approach to directly estimate capacity. The product of this approach is a new dataset named $\nu-$CLEAR. Rather than relying on theoretical arguments about the relative importance of nuclear activities, we "let the data speak," estimating the relative weights that should be placed on each activity as well as our proficiency scores. As such, a useful byproduct of this estimation was information about nuclear-related activities themselves.

The estimates we provide also have utility for policymakers. Nuclear proficiency plays a role in many policy considerations. As such, the quality of these estimates may have critical consequences for international outcomes. By providing a transparent measure that accounts for variation in observed indicators, this project provides another criteria for policymakers to use when considering issues related to nuclear security and proliferation.

While this project represents an improvement over existing measurement approaches, a number of future extensions are also apparent. One attractive extension of the current model is to explicitly account for dynamics in the estimation procedure. By incorporating dynamics into the model, shifts in each state's estimated proficiency could be dampened by information about their proficiency in previous periods. Such an extension would ameliorate the "lumpiness" that can be seen along the vertical axis in Figure 4. Another useful extension would involve explicitly modeling potential dependence among activities. We believe both of these are important directions that should be considered in future studies.

Finally, for these scores to be useful, it is important that they be robust to future data collection efforts and theoretical advancements. As such, an important component of this project is developing a set of tools to allow scholars to utilize the item-response approach themselves. With this, as new data arise or theoretical advancements point to new indicators of nuclear proficiency, other scholars may use our approach to improve upon the measures presented here. To do this, we provide user-friendly code to estimate new scores.

References

Abraham, Itty. 2006. "The ambivalence of nuclear histories." *Osiris* 21(1):49–65. 3

Ashraf, Quamrul and Oded Galor. 2013. "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development." *American Economic Review* 103(1):1–46. 23

Bailey, Michael, Anton Strezhnev and Erik Voeten. 2013. "Estimating dynamic state preferences from United Nations voting data.". 2

Bas, Muhammet and Andrew Coe. 2016. "A Dynamic Theory of Nuclear Proliferation and Preventive War." *International Organization* . 1

Benson, Brett and Quan Wen. 2011. A Bargaining Model of Nuclear Weapons: Development and Disarmament. In *Causes and Consequences of Nuclear Proliferation*, ed. Matthew Kroenig Robert Rauchhaus and Erik Gartzke. New York: Routledge Press. 1

Benson, Brett V and Joshua D Clinton. 2014. "Assessing the Variation of Formal Military Alliances." *Journal of Conflict Resolution* pp. 1–33. 2, 3

Bleek, Philipp C. 2010. Why do States Proliferate? Quantitative Analysis of the Exploration, Pursuit, and Acquisition of Nuclear Weapons. In *Forecasting Nuclear Proliferation in the 21st Century: Volume 1, the Role of Theory*, ed. William C Potter and Gaukhar Mukhatzhanova. Stanford: Stanford University Press. 11

Carpenter, Bob, A Gelman and M Hoffman. N.d. "Stan: a probabilistic programming language." . Forthcoming. 12

Chadefaux, Thomas. 2011. "Bargaining over power: when do shifts in power lead to war?" *International Theory* 3(02):228–253. 17

Coe, Andrew J and Jane Vaynman. 2015. "Collusion and the Nuclear Nonproliferation Regime." *The Journal of Politics* 77(4):983–997. 17

Debs, Alexandre and Nuno P Monteiro. 2014. "Known unknowns: Power shifts, uncertainty, and war." *International Organization* 68(1):1–31. 1, 17

Fuhrmann, Matthew. 2012*a*. *Atomic Assistance: How 'Atoms for Peace' Programs Cause Nuclear Insecurity*. Cornell University Press. 3

Fuhrmann, Matthew. 2012*b*. "Splitting Atoms: Why do countries build nuclear power plants?" *International Interactions* 38(1):29–57. 1, 6

Fuhrmann, Matthew and Benjamin Tkach. 2015. "Almost nuclear: Introducing the Nuclear Latency dataset." *Conflict Management and Peace Science* 32(4):443–461. 1, 4, 8, 11, 25, 26

Fuhrmann, Matthew and Sarah E Kreps. 2010. "Targeting nuclear programs in war and peace: A quantitative empirical analysis, 1941–2000." *Journal of Conflict Resolution* . 25

Horowitz, Michael C. 2013. "Nuclear Power and Militarized Conflict: Is There a Link?" *The Nuclear Renaissance and International Security* pp. 288–312. 1

Hymans, Jacques EC. 2006. *The psychology of nuclear proliferation: Identity, emotions and foreign policy*. Cambridge University Press. 1, 17

Jackman, Simon. 2008. Measurement. In *The Oxford Handbook of Political Methodology*, ed. Henry E. Brady Janet M. Box-Steffensmeier and David Collier. New York: Oxford UP. 1, 4, 27

Jo, Dong-Joon and Erik Gartzke. 2007. "Determinants of nuclear weapons proliferation." *Journal of Conflict Resolution* 51(1):167–194. 1, 4, 5, 6, 10, 11, 14, 15, 19, 23, 24, 25

Kroenig, Matthew. 2010. *Exporting the Bomb: Technology Transfer and the Spread of Nuclear Weapons*. Cornell University Press. 3

Martin, Andrew D and Kevin M Quinn. 2002. "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis* 10(2):134–153. 2

Mattiacci, Eleonora and Benjamin T Jones. 2016. "(Nuclear) change of plans: What explains nuclear reversals?" *International Interactions* 42(3):530–558. 1

Mehta, Rupal N. and Rachel E. Whitlark. 2016. "Benefits and burdens of nuclear latency." *Manuscript, University of Nebraska* . 1

Meyer, Stephen M. 1986. *The dynamics of nuclear proliferation*. University of Chicago Press. 4

Montgomery, Alexander H and Scott D Sagan. 2009. "The perils of predicting proliferation." *Journal of Conflict Resolution* 53(2):302–328. 25

Paul, Thazha Varkey. 2000. *Power versus prudence: Why nations forgo nuclear weapons*. Vol. 2 McGill-Queen's Press-MQUP. 25

Pemstein, Daniel, Stephen A Meserve and James Melton. 2010. "Democratic compromise: A latent variable analysis of ten measures of regime type." *Political Analysis* pp. 1–24. 2

Poole, Keith T and Howard Rosenthal. 1991. "Patterns of congressional voting." *American Journal of Political Science* pp. 228–278. 2

Rasch, Georg. 1960. *Probabilistic models for some intelligence and attainment tests.* ERIC. 5

Rubin, Donald B. 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81 John Wiley & Sons. 23

Sagan, Scott D. 1996. "Why do states build nuclear weapons? Three models in search of a bomb." *International Security* 21(3):54–86. 1, 3

Sagan, Scott D. 2010. Nuclear latency and nuclear proliferation. In *Forecasting Nuclear Proliferation in the 21st Century*, ed. Gaukhar Mukhatzhanova William Potter. Stanford: Stanford Security Studies. 1, 9

Sagan, Scott D. 2011. "The causes of nuclear weapons proliferation." *Annual Review of Political Science* 14:225–244. 2

Schnakenberg, Keith E and Christopher J Fariss. 2014. "Dynamic patterns of human rights practices." *Political Science Research and Methods* 2(01):1–31. 2, 23

Spaniel, William. 2015. "Arms Negotiation, War Exhaustion, and the Credibility of Preventive War." *International Interactions* 41(5):832–856. 1, 17

Stoll, J Richard. 1996. "World production of latent nuclear capacity." *URL: http://es. rice. edu* 80. 4

Treier, Shawn and Simon Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52(1):201–217. 1

Volpe, Tristan A. 2017. "Atomic Leverage: Compellence with Nuclear Latency." *Security Studies* 26(3):517–544. 1, 9