Survivability not Superiority: A Critique of Kroenig (2013)

Dana Higgins[*], Connor Huff[†], Anton Strezhnev[‡]

Draft

May 1, 2013

Abstract

Kroenig (2013*b*) finds that in crises between nuclear-armed states, countries possessing nuclear arsenals larger than those of their opponents tend to be victorious. After correcting for coding errors in the dataset and for finite-sample bias in clustered standard error estimates, we show that the original conclusion is too over-confident. We further demonstrate that the association between nuclear superiority and crisis victory is extremely sensitive to the author's variable coding decisions and model specifications. Using a new method for evaluating coding uncertainty, we find that under reasonable alternative choices, the nuclear superiority finding is no longer statistically significant at all conventional significance levels. We find instead that the possession of an assured nuclear second-strike capability is consistently and robustly associated with positive crisis outcomes among nuclear states. Survivability, rather than superiority, appears to be the element of a state's nuclear arsenal that has the most significant bearing on its ability to win nuclear crises. However, we urge caution in drawing any strong conclusions from the Kroenig dataset due to sample size and model dependence issues.

[*] Harvard University, Department of Government, danahiggins@fas.harvard.edu

[†] Harvard University, Department of Government, cdezzanihuff@fas.harvard.edu

[‡] Corresponding Author, Harvard University, Department of Government astrezhnev@fas.harvard.edu

1 Introduction

Does nuclear superiority influence the outcomes of international crises between nuclear tates? That is, does the relative size of a state's nuclear arsenal provide it with greater resolve in inter-state contests. Over the years the factors that shape these crises has been the subject of intense scholarly debate. This debate was recently reinvigorated by a pair of articles published in the January 2013 issue of *International Organization*. Kroenig (2013*b*) finds that nuclear states possessing superior arsenals are more likely to win international crises while Sechser & Fuhrmann (2013*b*) finds that nuclear superiority does not make states more likely to successful make compellent threats. This paper replicates Kroenig (2013*b*) and tests the sensitivity of the original findings.

Kroenig models crises between two nuclear-armed states as competitions in risk taking, where states unwilling to concede the bargaining contest risk the possibility of escalation to nuclear exchange. Because nuclear superior states are assumed to be more likely to win in a nuclear exchange, they face fewer costs in the event of escalation, are more willing to run the risk of conflict, and are therefore more likely to succeed in disputes with other nuclear states. Kroenig tests this hypothesis on a dataset of 26 dyadic disputes during the period $1946 - 2001$. Using a probit regression model to control for potential confounding covariates that would also affect a state's crisis resolve or correlate with nuclear superiority (i.e. conventional superiority, second strike capability, crisis stakes), the article suggests that nuclear superiority has a positive and statistically significant effect on the probability of crisis victory.

In this paper we find that the finding in Kroenig is extremely unrobust and does not hold under reasonable alternative modeling and coding decisions.[1] The original findings substantially understate estimation uncertainty and inferences about nuclear superiority are extremely dependent on both variable coding decisions and parametric model choice. We do not find sufficient evidence to support an association between nuclear superiority and nuclear bargaining outcomes. Instead, we find that possession of an assured second strike capability is consistently and significantly associated with crisis victory. While we are cautious to note that it is very difficult to draw any strong conclusions from such a small dataset, the evidence demonstrates that survivability

[1] Note that our paper does not directly critique the game-theoretic model proposed in the original paper. Rather, we demonstrate how it is extremely difficult (if not impossible) to make inferences based on the available data because nuclear crises are such rare events characterized by high levels of uncertainty. Indeed, our tentative finding that possession of a survivable rather than a superior arsenal follows from a variation of the Kroenig model.

rather than superiority is the component of nuclear posture that has the most consistent impact on crisis victory.

Our paper is structured as follows. Section two demonstrates that the core conclusion of Kroenig (2013*b*) no longer holds once we correctly compute uncertainty in the estimated effect of nuclear superiority. We find that nuclear superiority is no longer significant at the 5% level without making a single alternation to any of the substantive coding decisions made in the original paper. Sections three and four discuss a number of questionable variable coding decisions used in the Kroenig dataset and propose several revisions. Using a novel simulation method we show that only a minimal level of confidence in our revisions is necessary to further reject the conclusion that nuclear superiority affects victory at lower significance levels. Section five demonstrates that the original finding is also highly dependent on arbitrary modeling assumptions. Because the nuclear superior and inferior cases rarely overlap on relevant covariates, inferences are heavily dependent on the chosen parametric model. Section six demonstrates our finding that the possession of an assured second strike capability is consistently and significantly associated with crisis victory and illustrates its robustness across multiple specifications. We conclude by discussing the implications of our results for debates over nuclear force posture.

2 Correcting original standard error estimates

The Kroenig dataset contains observations at the directed-dyad level. That is, each pairing of countries involved in a nuclear crisis is included in the dataset twice. In each case, the independent variables, i.e. nuclear superiority, are measured for one of the two sides of the dyad. While this approach is a common practice for time-series cross-sectional studies of international interactions where independent variables are measured at the state level, it leads to model misspecification (Hoff & Ward 2004, Gartzke & Gleditsch 2008). Contrary to standard regression assumption of independence, the outcome of one directed-dyad observation is strongly correlated with the outcome of its converse. For example, if the United States achieves victory over the USSR in one crisis, by definition the USSR does not defeat the United States. Sechser & Fuhrmann (2013*a*) correctly notes that, absent correction, violation of the independence assumption will lead to downwardly biased standard errors. Failing to account for this dependence effectively inflates the "true" number of observations in the data.

Kroenig estimates and reports cluster-robust standard errors (CRSE) in order to account for dependence between directed-dyads. CRSEs are a class of "robust" variance estimators for regression coefficients that are consistent and unbiased even in the presence of heteroskedasticity and intra-cluster correlation.[2] In Kroenig, each cluster consists of pair of directed-dyad observations for a given cluster-dyad. So while the dataset consists of 52 reported observations, those observations are grouped into a total of 26 clusters.

We replicate Kroenig (2013*b*) and find, as does Sechser & Fuhrmann (2013*a*), that the clustered standard errors reported in the original paper's regressions are significantly underestimated due to an improperly coded cluster identifier. The corrupted variable only identifies 17 clusters rather than the correct 26. After recoding the cluster ID variable and re-estimating the fully specified model (Model 2), we obtain a standard error on the coefficient for nuclear superiority that is around 43% larger than reported.[3] With this correction alone, the effect of nuclear superiority remains statistically significant, but only at the 5% level.

However, even these standard error estimates are likely to be too small. Cluster-robust standard errors are consistent and unbiased as the number of clusters goes to infinity. In small samples, with very few clusters, the estimator will tend to under-estimate the true variance. The standard correction for this bias is to multiply the estimates by $\frac{M}{M-1}$ where M is the total number of clusters. This is the default correction used by the STATA statistical software for CRSEs in non-linear models, which we denote *HC1*.[4] Although it is commonly used, this correction is anti-conservative.

MacKinnon & White (1985) shows that for typical Huber-White heteroskedasticity-consistent variance estimators that this simple *HC1* adjustment is an inadequate correction for small sample bias and presents a set of alternative estimators with substantially improved finite sample properties. Bell & McCaffrey (2002) extends the MacKinnon and White bias-reduced estimator to the case of CRSEs and demonstrates that this modification more accurately estimates the true coefficient variance when the number of groups or clusters is small. We denote this correction *HC2*.

[2] CRSE estimators are generally derived from the original heteroskedasticity-consistent "sandwich" estimator developed by Huber (1967) and White (1980). Arellano (1987) extends the Huber-White estimator to the case of clustered data.

[3] We focus our replication primarily on the fully-specified model since by the author's own account, the relationship between nuclear superiority and crisis outcome is likely to be confounded by other relevant covariates.

[4] For linear models, STATA instead uses $\left(\frac{M}{M-1}\right)\left(\frac{N}{N-K}\right)$ where N is the number of observations and K is the number of parameters in the model. This corrects both for small cluster bias and small sample bias (low N).

Imbens & Kolesar (2012) strongly recommend that applied researchers adopt this method even in reasonably sized samples due to the extraordinarily poor properties of the more commonly used estimators when the number of clusters is small. Indeed, Angrist & Lavy (2009) finds that using a variant of *HC2* instead of *HC1* increased cluster-robust standard error estimates by roughly 10 to 40 percent in a datset of 30 to 40 clusters. Monte Carlo simulations by Cameron et al. (2008) further demonstrate that *HC2*-corrected standard errors lead to rejection rates that are much closer to the desired level.[5] Simply put, typical cluster-robust standard error estimators will still lead to over-rejection in small samples. Implementing further small sample corrections will improve the reliability of our inferences about the effect of nuclear superiority.[6]

Figure 1 compares the coefficient estimates and 95% confidence intervals for nuclear superiority reported in the original paper with our re-estimated values and confidence intervals using the *HC1* and *HC2* corrections. Using the improved *HC2* estimator further increases the estimated standard error and the 95% confidence interval overlaps 0. The p-value of the significance test for whether the effect of nuclear superiority on crisis victory is distinguishable from 0 equals 0.7198.[7] After we correct for coding errors and use an improved standard error estimator, the relationship between nuclear superiority and victory is no longer statistically significant at the 5% threshold used by Kroenig to assess significance. Without a single alteration to the substantive coding decisions made in the original paper, the core conclusion of Kroenig (2013*b*) no longer holds once we correctly compute the uncertainty in the estimated effect of nuclear superiority.

Moreover, because the estimated level of uncertainty is much greater once we make the proper corrections, the nuclear superiority finding fails many of the robustness checks reported in the original paper. Kroenig (2013*b*) reports that the statistically significant relationship between superiority and crisis victory is not driven by the inclusion of any specific crisis. Following the approach taken in the original article, we sequentially remove each crisis from the dataset and re-estimate the model.[8] Figure 2 reports the estimated probit coefficients and 95% confidence

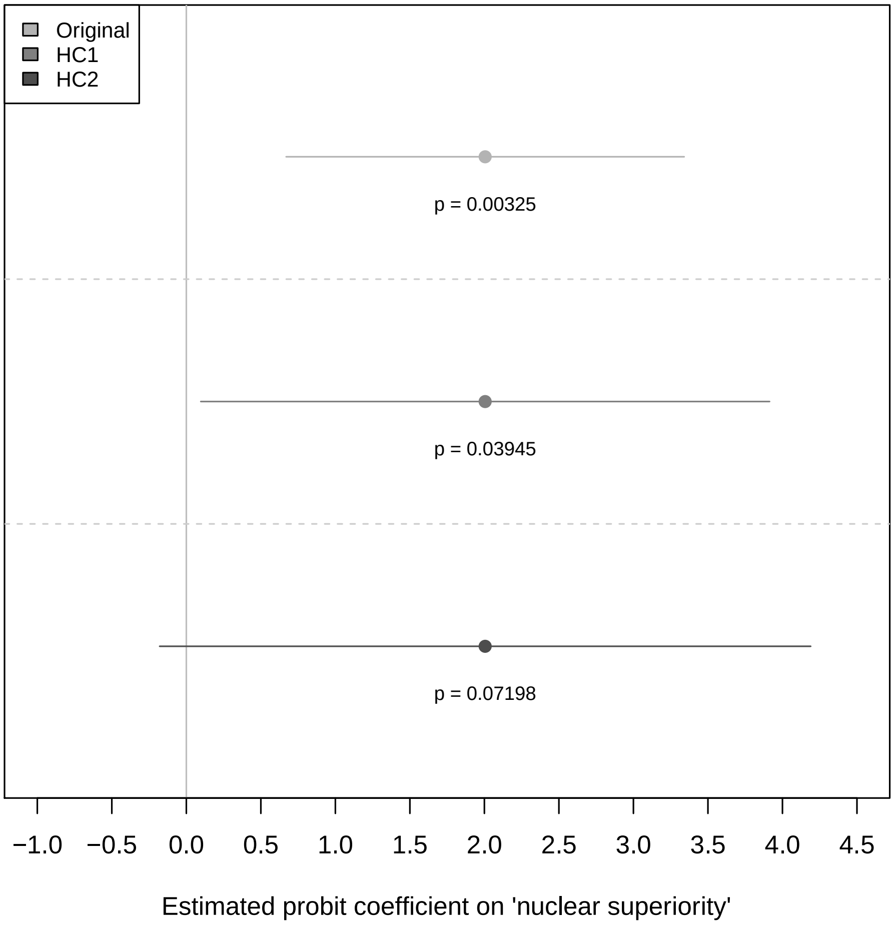
[5] It is worth noting that even *HC2*-corrected errors still underestimate the true level of estimation uncertainty. (Cameron et al. 2008) recommend using bootstrap methods to improve accuracy. However, bootstrap methods break down on the Kroenig dataset since the use of binary regressors can often lead to perfect separation in GLM estimates on some bootstrapped samples.

[6] We present the exact formulae for *HC1* and *HC2*-corrected CRSEs in Appendix A.

[7] Using the small-sample variant HC1 correction $\left(\frac{M}{M-1}\right)\left(\frac{N}{N-K}\right)$ also leads us to reject at the 5% level with $p = .06164$

[8] We do not report the results when we omit the "Able Archer" crisis since the GLM estimator failed to properly converge, a consequence of the small dataset and high collinearity among the regressors.

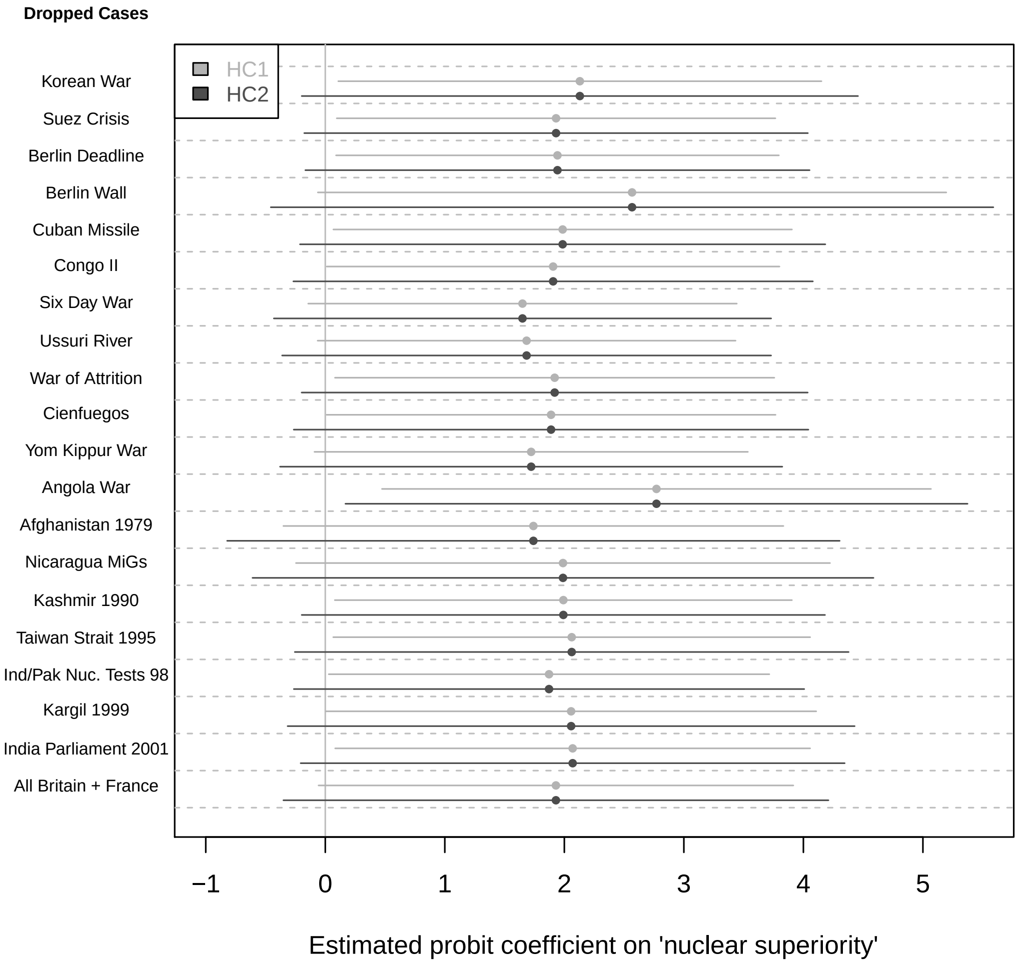
Figure 1: Estimated probit coefficients for 'nuclear superiority' variable - Original Kroenig (2013) dataset - Model 2



Note: Lines denote 95% confidence intervals.

intervals for each of these models. We find that, even when we use the most anti-conservative finite-sample correction *HC1*, the original superiority finding is not robust to removal of many of the crises, including the Afghanistan War (1979) and the Nicaragua MiGs crisis. Using the improved *HC2* finite sample correction leads to even larger confidence intervals for each of the models. The results only hold at the 95% level when we selectively remove the Angola War crisis. We also replicate another robustness check suggested by the author and re-estimate the model on a dataset that removes observations containing the U.K. and France. Because these states are allied with the United States in the post-WWII era, it is inappropriate to treat their involvement in crises and nuclear arsenals separately from the United States. Sechser & Fuhrmann (2013*a*) also notes this problem with the original dataset. Dropping these cases makes the coefficient on nuclear superiority statistically insignificant regardless of whether the *HC1* or *HC2* correction is used.

Figure 2: Estimated probit coefficients for 'nuclear superiority' variable - Kroenig (2013) robustness checks - Model 2



Note: Lines denote 95% confidence intervals. Results of omitting the "Able Archer" crisis not shown as the estimator failed to converge due to separation.

3 Revising coding decisions

In addition to substantially understating estimation uncertainty, Kroenig (2013*b*) makes a number of contestable decisions in coding variables used in the regression dataset. The chosen values of the dependent variable - crisis outcome - rely on subjective judgements about the historical interpretation of crisis events. Additionally, because of the extensive secrecy surrounding states' nuclear weapons programs, exact arsenal size estimates are inexact and it is often unclear which state in a dyad is the "nuclear superior" one. This section proposes and explains a set of reasonable revisions to the author's original dataset.

Some of the cases in the dataset do not qualify as nuclear crises as they either do not involve direct action by two nuclear-capable actors. We suggest excluding three crisis-dyads from the dataset. First, France had no arsenal during its involvement in the Berlin Wall crisis. A successful nuclear test was conducted in 1960, but France possessed no nuclear weapons until 1964 after the crisis was over. The original dataset even codes the size of the arsenal as 0, meaning that France was not capable of using nuclear weapons. Therefore dyads containing France in 1961 are inappropriately included in the data.

Additionally, during the Six Day War, Israel had no confirmed usable nuclear weapon. It is extremely difficult to get accurate estimates of the size of Israel's nuclear arsenal given that Israel continues to officially deny its existence. Kroenig relies on the Federation of American Scientists which states "it is widely reported" that Israel had installed nuclear warheads on two missiles on the eve of the conflict." However, according to reports from the Central Intelligence Agency, Israel did not have nuclear weapons during the Six Day War but developed them later the same year (Aftergood & Kristensen 2007). In a declassified record of a National Security Council meeting held on May 24, 1967, the Director of Central Intelligence "was quite positive in stating there were no nuclear weapons in the area" (National Security Archive 2006). In the same record, U.S. Ambassador to the UN, Arthur Goldberg confirmed that he was "less certain about Israeli superiority." The Bulletin of Atomic Scientists confirms that even today it is difficult to assess whether or not there were changes in Israel's nuclear arsenal at the time (Norris & Kristensen 2009). Kroenig codes the Six Day War as a crisis between Israel and the Soviet Union. But it is difficult to imagine that Israel could threaten the U.S.S.R. with its unknown nuclear arsenal or that the U.S.S.R. would even consider the possibility of the crisis escalating to nuclear use against

Israel. Therefore, Israel should only be included in the dataset for post-1967 crises.

Finally, we recommend removing observations for the Congo II crisis since the United States was not a true participant. The original dataset treats Congo II as a crisis between the U.S. and the U.S.S.R. However, U.S. involvement was limited to assisting Belgian forces with a rescue mission, a mission which initiated the crisis and prompted Soviet disapproval but had no direct interaction with the Soviet Union or its forces. Some descriptions of the crisis describe the intervention as uniquely Belgian with no mention of United States or British support (Clarke 1968, 18-20). Other sources indicate a stated willingness to provide the Belgian intervention with transport and communications, but no troops or more direct involvement (Gleijeses 1994, 216). Indeed, Sechser & Fuhrmann (2013*b*) highlights the sparse participation of the U.S. in the Congo II crisis as a flaw in in analysing every dyadic ICB crisis interaction. The ICB dyadic dataset attempts to include every crisis participant regardless of their level of involvement.

In addition to these inappropriate case inclusions, the dataset contains over-confident assessments of nuclear superiority. Kroenig (2013*b*) finds reasonably accurate estimates of P5 arsenal sizes from National Resources Defense Council (NRDC) and Federation of American Scientists (FAS) data. However, for the non-NPT nuclear powers, Israel, India, and Pakistan, estimates of nuclear superiority – much less specific arsenal size – are highly uncertain; most estimates offer at best a wide range. Current estimates, for instance, of Israel's stockpile range anywhere from 70 to 400 warheads (Aftergood & Kristensen 2007). Conveniently, Israel is only coded as being involved in crises with the Soviet Union, which most certainly is the nuclear superior actor.

However, uncertainty over Indian and Pakistani arsenal sizes has a meaningful impact on the results since these states confront one another. These estimates are complicated by government secrecy and the use of using approximate measures based on each country's supply of nuclear material. Estimates of India's arsenal size over the past decade range from effectively zero weapons (Tellis 2001) to as many as 90 (Cirincione et al. 2002). Estimates of Pakistan's arsenal size range from a low of 24 to a high of 90 weapons (Norris & Kristensen 2009). These ranges tend to overlap in the relevant time periods making it difficult to determine which actor is the "superior" one. The author attempts to account for this uncertainty by estimating arsenal size as the mean of a variety of estimates (using the median of a ranged estimate), but does not disclose from what sources these estimates are calculated. As an average, the estimates are highly dependent on the selection of source material since an outlier could greatly alter the mean arsenal size.

Rather than making assertions about the exact arsenal size, we conduct a set of robustness checks by re-estimating the regression model under the three potential superiority combinations (India superior, Pakistan superior, or no clear superiority). We replicate the simple chi-squared test reported in the original paper and find the significance of the bivariate relationship between victory and superiority holds only under the assumption that India is superior. While this assumption might be reasonable for the early 1990s crises (ex. Kargil III), it is less reasonable as Pakistan accelerated its arsenal development around the turn of the century. U.S. intelligence analyses in the wake of the 1998 nuclear tests suggest that the size of India's arsenal at the time was significantly overestimated while Pakistan's was underestimated. According to declassified U.S. Defense Department documents, Pakistan had weaponized most of its fissile material stockpile while India had not, making prior assessments based on known nuclear materials possesion inaccurate (Albright 2000). Indeed, a number of U.S. military and intelligence officials at the time strongly questioned the long-standing assumption of Indian superiority (Windrem & Kupperman 2000). Even if Pakistan remained quantitatively inferior to India by a few warheads, it is not reasonable to treat a minor warhead difference as a clear indicator of which state is nuclear-dominant. Due to the wide ranges provided by most estimates and the high uncertainty even today of nuclear strength at the time of crisis, we suggest opting for the more conservative option and coding no clear nuclear superiority between the two states for the 1999 and 2001 crises in the dataset.[9]

Moreover, the dataset contains a number of questionable classifications of crisis outcomes variable. The author relies primarily on the International Crisis Behavior Project's OUTCOM variable to code the outcome of each nuclear crisis.[10] The ICB coding is coarsened into a binary variable where an ICB victory is coded as 1 (success) and an ICB compromise, stalemate, or loss is a 0 (failure). Although Kroenig (2013*a*) notes that the ICB dataset is "well-worn, has been accepted by the scholarly community, and has stood the test of time," the original paper makes several alterations to the ICB coding of outcomes. We suggest reverting two of these changes back to their original coding.[11]

[9] Although the author presents results for a continuous measure of nuclear superiority, there is simply too much uncertainty over the size of nuclear arsenals to make anything but a binary indicator useful. As we show here, even identifying *which* state is superior at a given point in time is extremely challenging let alone identifying the precise ratio of superiority.

[10] The variable is coded as victory (1), compromise (2), stalemate (3), or defeat (4).

[11] One alteration that we concur with is the exclusion of the Anglo-Icelandic Cod War of 1973 as there is no

The author recodes the Ussuri River crisis in 1969, a border dispute between China and the Soviet Union primarily over two islands in the river, as a victory for the Soviet Union. The author's stated justification is limited to Henry Kissinger's personal memoirs. However, according to most scholars, including Yang Kuisong whom the author cites, the conflict was a stalemate.[12]. Yang argues that Mao had domestic goals for instigating the border conflict over Zhenbao island on the Ussuri River. Though the conflict eventually escalated beyond Mao's intentions, his domestic goals were arguably achieved, making a Soviet success and Chinese defeat an unreasonable coding decision (Yang 2000). Indeed, after the September 1969 peace agreement, Chinese troops were permitted to remain on the disputed Zhenbao island - hardly a Soviet victory (Ryabushkin 2007).

Further, the original Kroenig dataset treats the Korean II crisis as a stalemate citing Gaddis (2005). However, the original ICB coding of Soviet victory is more appropriate. The crisis involved United States forces crossing the 38th parallel triggering a crisis for the Soviet Union. The stated goal of the United States, as articulated by General MacArthur, was the unconditional surrender of North Korea, a goal that was not achieved (Appleman 1989). The objective of the Soviet Union was to prevent United States forces from entering Soviet territory or making significant advances in the north. This goal was fully accomplished when U.S. and South Korean forces were pushed back south of the 38th parallel in December.

Lastly, the Sino-U.S. Taiwan Strait crisis in 1995 is considered in both the Kroenig and the ICB datasets to be a victory for the United States, implying a defeat for China. The crisis occurred because China attempted to prevent pro-Taiwan movements from gaining influence in Taipei and to prevent Taiwan from officially declaring independence. To achieve this goal, Chinese forces deployed – officially described as an exercise – in the coastal area facing Taiwan. The forces had the potential to conduct a military operation. In response, the US deployed a carrier to deter a possible invasion. No war broke out and the PRC chose not to invade, but both nuclear powers ultimately achieved some amount of their goals. The goal of the United States, to avoid the outbreak of violence, was accomplished. However, this assumes that China had a real intention to invade Taiwan. Indeed, a number of Chinese objectives were arguably accomplished. The United States weakened its strong support of Taiwan and instead recognized China's reunification policy (Russell 2000). Furthermore, China prevented a significant rise in pro-Taiwanese sentiments and

convincing evidence of Soviet intervention on the part of Iceland

[12] Kroenig cites Yang (2000) as "Yang Gongsu" in the bibliography to the appendix

Table 1: Summary of proposed coding revisions

Case	**Revision**	**Justification**
France, 1961. Berlin Wall crisis	Remove from dataset	France did not have an arsenal until 1964
Israel, 1967. Six Day War	Remove from dataset	CIA reports suggest Israel developed a weapons capability *after* the war.
Congo II crisis, 1964	Remove from dataset	Minimal U.S. involvement
U.S.S.R., 1969. Ussuri River	Recode U.S.S.R. outcome as stalemate (original ICB coding)	Only citation for USSR victory is Kissinger's personal assessment. Historical consensus suggests both sides only partially achieved their objectives. Mao achieved domestic political goals and chinese troops were permitted to remain on Zhenbao island after the September settlement.
U.S.S.R., 1950. Korean War	Recode U.S.S.R. outcome as victory (original ICB coding)	U.S. failed to achieve North Korea's unconditional surrender. Soviet Union successfully prevented U.S. from advancing north.
U.S., 1995. Taiwan Strait	Recode U.S. outcome as stalemate	No evidence that U.S. presence deterred a potential attack. Chinese military exercises achieved the goal of deterring Taiwan from declaring independence.
India/Pakistan, 1999/2001	Recode neither as nuclear-superior	Uncertainty over arsenal estimates for both countries overlaps - no *clearly* superior actor

successfully deterred Taiwan from declaring independence. Assuming that China's primary goal was to invade, the crisis was a victory for the United States. However, this is a rather tenuous interpretation of the crisis and we recommend recoding the crisis as a draw.

The full set of coding revisions is summarized in Table 1. The following section illustrates the effect of these changes on the conclusions of Kroenig (2013*b*). Because there is epistemic uncertainty over which coding choices are the "correct," we want to know how much certainty in the original dataset is necessary to sustain the original conclusion. To do so, we present a simulation method similar to missing data imputation to propagate uncertainty over dependent and independent variables into a set of regression estimates. We show that only a minimal level of confidence in our revisions is necessary to reject the conclusion that nuclear superiority is positively associated with crisis victory.

4 Sensitivity analysis for coding uncertainty

Along with basic estimation uncertainty that inevitably arises whenever one attempts to estimate some unknown quantity there also exists some level of *measurement* uncertainty that scholars often ignore. In binary dependent variable regression models like logit and probit, mismeasurement of both the dependent and independent variables will yield biased and inconsistent measures of the quantities of interest (Hausman 2001). A core assumption of these models is that all variables are measured perfectly. However, social scientists typically work *imperfectly* measured data. This has significant consequences for inference. For example, Treier & Jackman (2008) illustrates how failing to account for measurement uncertainty in the discipline-standard Polity indicator of regime type leads to overconfident assessments of democracy's impact on conflict or economic development.

The typical method in political science for dealing with ambiguity in variable measurements and coding decisions is to conduct robustness checks and re-estimate the model under different assumptions. A finding that passes several robustness checks is considered unaffected by this form of measurement uncertainty. But when a finding is not robust to some alternative coding, it is often unclear how researchers should proceed. The author may then use subjective judgement to argue discard an alternative as too implausible. However, more involved sensitivity tests may provide some additional guidance in evaluating an author's justifications. If a finding is not robust to a particular coding decision, we want to know how much confidence we need in the alternative such that the original finding no longer holds. That is, how strong must the author's defense be.

We propose an easy to implement simulation method for conducting this sensitivity test. The approach is inspired by multiple imputation methods for missing data developed by Rubin (1987) and later in King et al. (2001). The intuition is that the disputed coding problem is analogous to the missing data problem. We assume that there exists some true value of a particular variable, but are unsure of the exact value. In the missing data context, these are the variables left unobserved due to non-response or lack of data availability.

Multiple imputation methods use statistical models to generate a distribution of predicted values for each missing data entry. Each missing value is imputed m times by drawing from the predicted distribution, generating m separate datasets. One can then estimate any quantity of interest (i.e. a regression coefficient) on each of the datasets and combine the estimates into a single result using the procedure described in Rubin (1987) and King et al. (2001).

Let K be the quantity of interest. The point estimate for K, $\bar{k}$, is the average of all of the point estimates k_i, ($i = \{1, \ldots, m\}$) obtained from each of the m simulated datasets.

$$\bar{k} = \frac{1}{m}\sum_{i=1}^{m} k_i$$

The variance of the estimator is:

$$Var(\bar{k}) = \frac{1}{m}\sum_{i=1}^{m} Var(k_i) + S_k^2(1 + 1/m)$$

Where $Var(k_i)$ is the estimated variance of k_i and $S_k^2 = \sum_{i=1}^{m}(q_i - \bar{q})^2/(m-1)$ or the sample variance of the m point estimates.

This multiple imputation approach can also be used as a way of including uncertainty about coding decisions into uncertainty about estimated parameters. We define the Sensitivity Analysis for Coding Uncertainty (SACU) algorithm as follows:

1. Choose m, the number of simulated datasets

2. For each disputed value or set of values, define a hypothetical distribution over possible coding decisions.

3. For each simulated dataset i in $1, \ldots, m$

    (a) Impute each disputed value by drawing a coding decision from the specified distribution.

    (b) Conduct the statistical procedure and obtain an estimate of the quantity of interest k_i.

4. Using the formulae defined above, combine the m estimates of the quantity of interest into a single point and interval estimate.

If we have prior knowledge of the level of uncertainty around a particular value, for example, an empirical distribution of judgements by different experts, then we can use this information to construct an imputation distribution for each "uncertain" value. This way, the known level of measurement uncertainty can propagate into the estimated coefficients and standard errors.

Conversely, we can define some arbitrary discrete distribution over proposed alternative codings in order to more precisely evaluate the sensitivity of the author's original dataset choices. For each value j, we choose some $\delta_j \in [0, 1]$ to be the subjective probability that the author's coding for that value is correct. The probability that we believe the alternative coding is correct is therefore $1 - \delta_j$. This yields a vector $\boldsymbol{\Delta}$ representing the proposed level of confidence in the author's coding. Conducting the simulation, we obtain a point estimate and uncertainty for our quantity of interest *given* a certain level of belief $\boldsymbol{\Delta}$. For simplicity's sake, we will set all δ_j to some single certainty level δ. We can then sweep across the space of possible δ values to identify the general degree of certainty in the author's coding decisions necessary to sustain a finding.

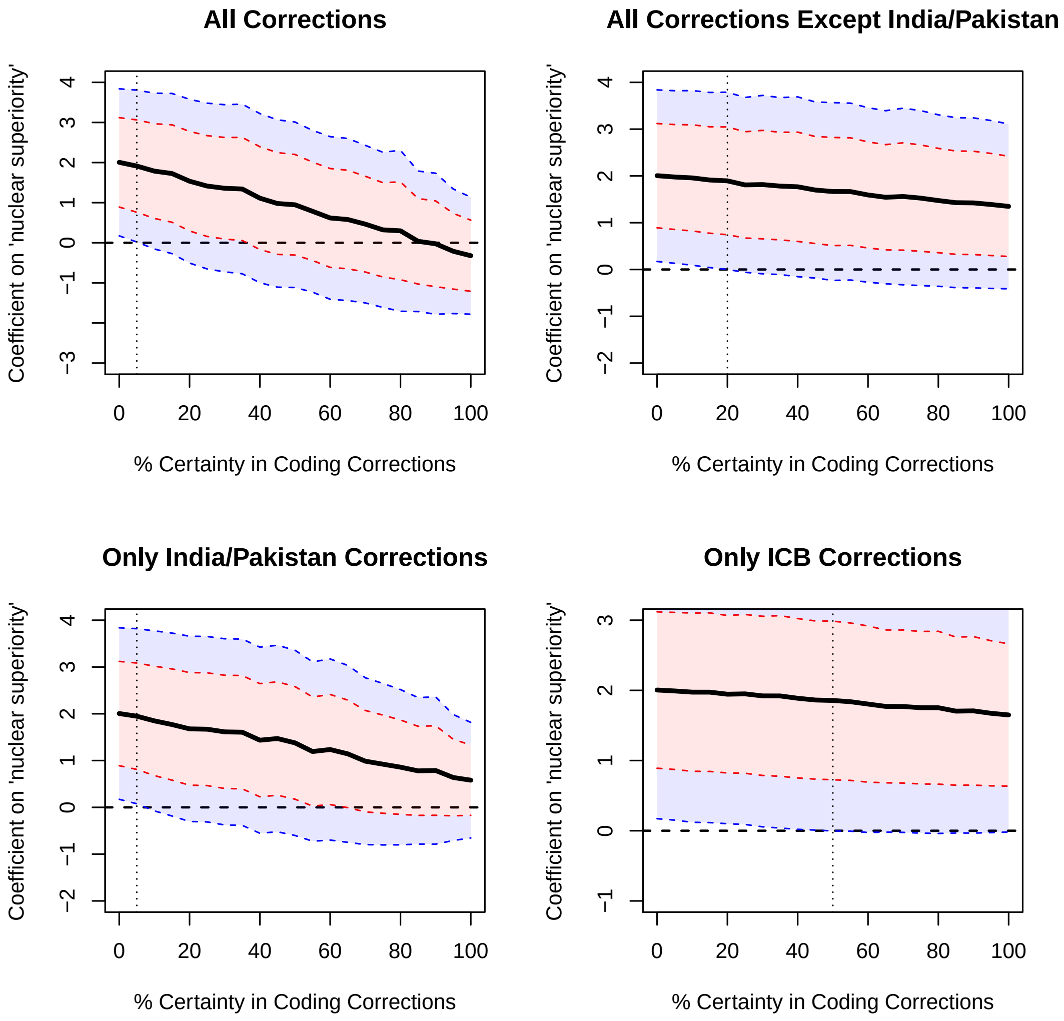
We apply this method to examine how sensitive the findings of Kroenig (2013*b*) are to a set of proposed coding revisions. We have already shown that the effect of nuclear superiority is not significant at the 5% threshold. Using a 10% threshold reduces the probability of a Type II error (failing to reject when the null hypothesis is false), but raises the probability of a Type I error (rejecting when the null hypothesis is true). The trade-off between Type I and II error rates is ultimately a subjective decision. Is it better to be overconfident and conclude that nuclear superiority leads to crisis victory when in reality it does not or to be overcautious and find that nuclear superiority does not matter when in truth it does?

For the purposes of this analysis we will grant the author the benefit of the doubt and use a more powerful but less conservative test that rejects when $p < .10$. However, even with this reduced threshold, we only need a small degree of uncertainty in the Kroenig (2013*b*) dataset in order to fail to reject the null.

Figure 3 plots the estimated probit coefficients on nuclear superiority and confidence intervals for each degree of certainty in our alternative codings.[13] We conduct the sensitivity analysis using the full set of proposed corrections along with three subsets: all corrections except for the changes to India and Pakistan's superiority status, only changes to India and Pakistan's superiority status, and only the two reversions to the original ICB outcome codings (Sino-Soviet Border Conflict and Korean War).

[13] We run the simulations for $\delta = 0$ to $\delta = 1$ in .05 unit increments. For each simulation we create $m = 200$ datasets and estimate HC2-corrected cluster-robust variance-covariance matrices.

Figure 3: Sensitivity analysis for coding uncertainty in Kroenig (2013) - Model 2



Note: Shaded areas denote 66% and 90% confidence intervals respectively. Level of coding certainty indicates the probability that each independent correction will be selected in a simulated dataset. 200 simulated datasets are created for every 5% certainty increment. HC2 cluster-robust variance matrices estimated for each iteration. Vertical dotted lines denote the minimum level of certainty in corrections at which we fail to reject the null hypothesis of 'no nuclear superiority effect' at the 10% level.

The results show that we only need around a 5% level of belief in our collective set of revisions in order to find the effect of nuclear superiority non-significant at the 10% level. In other words, with a minimal level of certainty in all of the proposed revisions we still cannot support the finding in Kroenig (2013*b*). Even if we were to believe that the superiority variable is coded perfectly and assign a probability of 0 to our proposed India/Pakistan correction, we only need about 20% certainty in the remaining six corrections to reject the original conclusion. Given how problematic many of the original case and outcome choices are, this is a very low bar.

Suppose that we *only* consider our two suggested reversions to the original ICB outcome coding in the Korean War and Ussuri River crises. The simulation demonstrates that only a roughly 50% level of certainty in these two codings alone is necessary to conclude that nuclear superiority is not statistically significant. That is, if we give the original ICB coding and the author's revisions *equal weight*, then we still cannot conclude at the 10% level that nuclear superiority has a statistically non-zero association with crisis victory. Given the ICB's longevity and relative acceptance among international relations scholars, 50% certainty in the ICB outcome codings relative to the changes in Kroenig (2013*b*) is definitely not unreasonable. The Kroenig (2013*b*) findings are not merely unrobust, they rest on an indefensibly high level of confidence in the correctness of the dataset. [14]

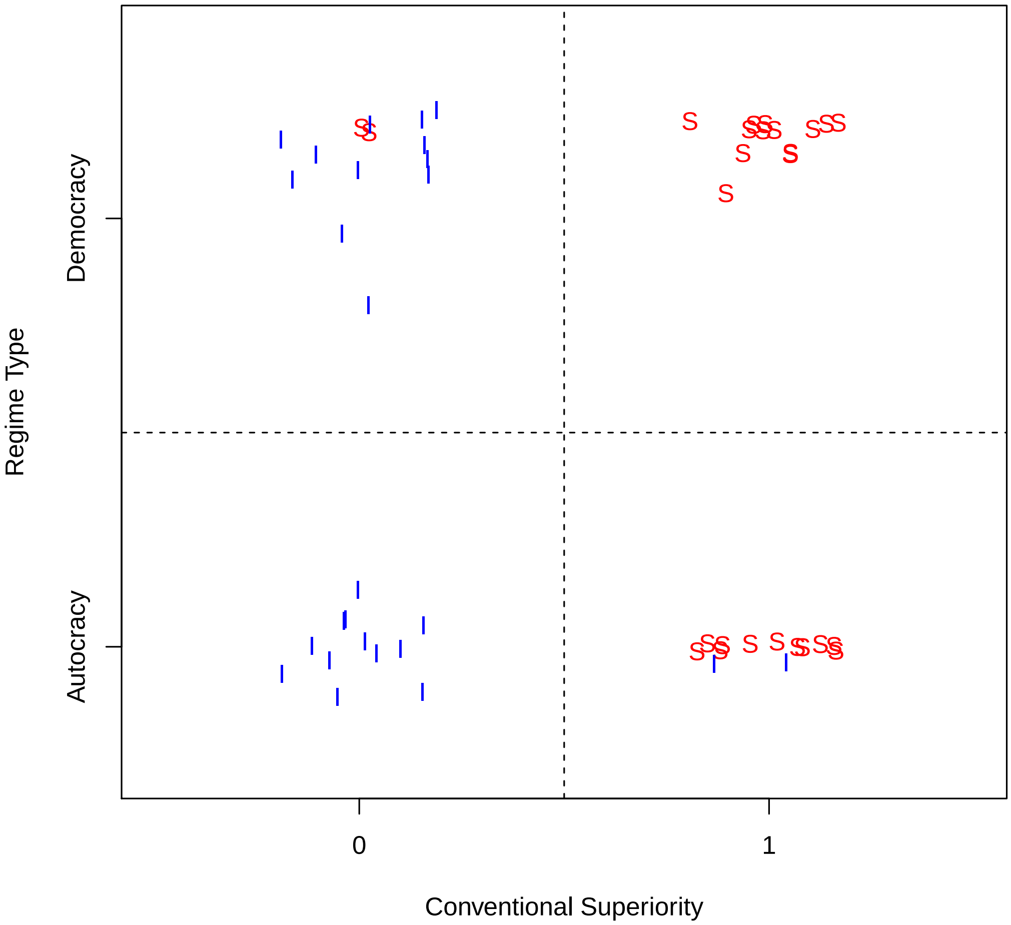
5 Model dependence

Regardless of whether the original original coding decisions are perfectly correct, the finding that nuclear superiority has a positive and statistically significant effect on crisis victory remains highly dependent on implicit and undefended modeling assumptions.

Identifying the causal effect of some variable on another requires comparing an observed factual observation to some unobserved *counterfactual*. The ideal counterfactual is some other observation that shares characteristics with the factual case but differs on the variable of interest. For some observations in the dataset, there may not be a suitable counterfactual. In this dataset, there is very little overlap between nuclear superior and nuclear inferior cases on relevant covariates. Typical practice in the social sciences is to use a parametric model (e.g. linear, logit, probit) to control for confounders and *estimate* counterfactuals where no overlap exists. However, as King & Zeng (2006) and King & Zeng (2007) illustrate, when there is little overlap in the covariate *support* of groups that differ in the causal variable of interest (in this case, the nuclear superior and nuclear inferior groups), inferences will be heavily dependent on the chosen parametric model. This is because counterfactuals must be extrapolated far away from the observed data. These are *extreme* counterfactuals in the sense that we do not observe any remotely analogous cases in the data. All we have is the model. Such dependence is problematic since scholars never know

[14] It is valuable to note that statistical significance is only part of the story and is typically over-emphasized by scholars in lieu of more substantive quantities. The important conclusion to draw from this discussion is simply that the level of uncertainty in the given data is too high to make any justifiable conclusions about nuclear superiority.

Figure 4: Difference in the support of Nuclear Superior (S) and Nuclear Inferior (I) groups



the "true" model or the correct functional relationship between the independent and dependent variables. Estimated causal effects are subject to extrapolation bias and there is no clear means of adjudicating between models - there simply is not enough data.

This is a substantial problem for Kroenig (2013*b*). Figure 4 illustrates the extremely minimal overlap between the support of the nuclear superior group and the support of the nuclear inferior group. Conventional superiority perfectly predicts nuclear superiority in all but two dyadic observations.[15] When combined with regime type, conventional superiority perfectly predicts nuclear superiority. The absence of a reasonable counterfactual illustrates that any causal inference made about nuclear superiority will require significant extrapolation and risk bias. The data contains

[15] These observations are the Angola War and the Afghanistan War, both involving the U.S. and the U.S.S.R.

zero nuclear inferior but conventionally superior democracies that could be matched to the bulk of nuclear and conventionally superior democracies. They simply don't exist in the historical record.

Indeed, minor alterations to the model significantly affect the results. To illustrate, we consider an alternate specification of the conventional superiority variable. Because there is no real theoretical justification for treating conventional superiority as a continuous and linear predictor of the outcome, we propose an alternate, binary, indicator of conventional superiority.[16] This mirrors the coarsened coding of nuclear superiority used in the author's original model. We re-estimate the full model on the Kroenig dataset using this binary indicator instead of capability share. With HC1-corrected robust standard errors clustered on crisis-dyad, the effect of nuclear superiority is not statistically significant at the 5% level ($p = 0.383$).

Alternatively, we consider using an ordinal rather than a binary outcome in order to capture additional variation in the dependent variable. The decision to treat stalemate and compromise as equivalent to defeat in the original regression analysis is entirely haphazard, dictated more by modeling convenience rather than by theoretical insight. Indeed, compromise or stalemate suggests a level of bargaining success greater than outright concession and defeat. An equally justifiable method of representing the dependent variable is to to coarsen the ICB 4-level outcome for each crisis-dyad into three levels: victory, defeat and an intermediate category that incorporates both stalemate and compromise.[17] Then, instead of estimating a standard probit model, we use an ordinal or ordered probit model which generalizes the binary probit to discrete ranked outcomes. When we use only the original ICB outcome codings and estimate the model, we find that the effect of nuclear superiority is again not statistically significant at the 5% level.[18] The effect is insignificant at 10% when we introduce the remainder of our proposed data corrections.

6 Assured Second-Strike Capabilities and Crisis Outcomes

The nuclear superiority finding of Kroenig (2013*b*) is incredibly sensitive to model and coding choices. Based on the data provided, there is insufficient evidence to credibly justify an association

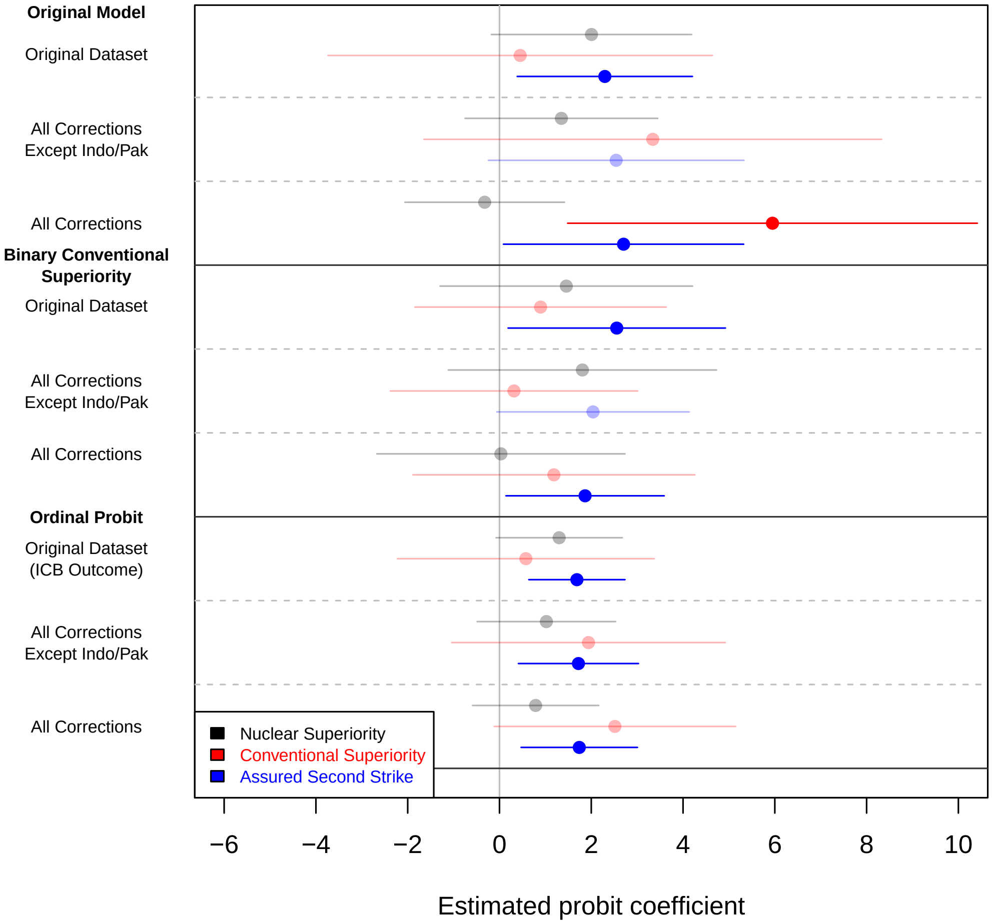
[16] The variable takes on 1 when a crisis-actor has $> 50\%$ of the dyadic share of capabilities and 0 when it has $< 50\%$.

[17] While the ordering of victory and defeat is clear, it is unclear whether a stalemate or a compromise is a superior outcome.

[18] We estimate HC1-corrected rather than HC2-corrected clustered standard errors for the ordered probit model since the numerical ML estimation routine we use utilizes Newton-Raphson rather than IRLS methods.

between superiority and crisis victory. Nuclear weapons policy is an incredibly serious topic and conclusions about the effect of nuclear-weapons postures have meaningful implications for states' nuclear force choices. As a result, we recommend erring on the side of caution rather than overconfidence. There simply is not enough data.

Figure 5: Estimated results for nuclear superiority, conventional superiority and assured second-strike capability on crisis outcome across multiple model specifications



Note: Lines denote 95% confidence intervals. HC2-corrected cluster-robust standard errors estimated for the binary probit models. HC1-corrected cluster-robust standard errors estimated for the ordered probit model.

But what, if any, relationship remains between nuclear weapons posture and crisis victory? Throughout our various regression analyses, we find that possession of an assured second-strike capability, that is, a difficult-to-destroy nuclear arsenal capable of retaliation in the event of a nuclear war, is consistently and significantly associated with crisis victory. While the positive effect of nuclear superiority remains highly dependent on particular model and coding choices, the positive effect of survivable nuclear arsenals is robust to most alternative specifications.

Figure 5 plots the estimated regression coefficients and 95% confidence intervals for nuclear superiority, conventional superiority and assured second-strike across the gamut of changes considered in this paper.[19] While we fail to reject the hypothesis of no effect for nuclear superiority at the 5% level for all considered specifications, we are able to reject the null of no second-strike effect at the 5% level for all but two specifications (where $.10 > p > .05$). Having an assured second strike capability appears to, on average, increase the probability that a state will achieve victory in a crisis. It is important to note that this finding is tentative; the sparseness of the data does not permit us to draw strong conclusions. However, the relationship between second-strike capabilities and crisis victory is *much more robust* than the association between superiority and success. If we must draw a conclusion from this dataset, it is that survivability and not superiority is the element of nuclear deterrence that matters more for crisis outcomes.

Indeed, the logic for why second-strike capability leads to more positive crisis outcomes follows from the author's original bargaining model. The model assumes that the cost of an accidental nuclear war imposed by one state on another is increasing in that state's arsenal size. Nuclear superior states lose less than nuclear inferior states if a crisis escalates to the point of an actual nuclear exchange and will therefore have more bargaining leverage. There is an implicit assumption that leaders view some nuclear wars as more or less costly than others. However, there is significant debate amongst nuclear strategists about how to weigh the devastation from a nuclear conflict and whether this assumption is in fact true. For leaders considering the risk of nuclear devastation, the fine distinctions between $13,000$ and $12,999$ nuclear weapons are perhaps not so salient. Kroenig cites the views of nuclear warfighters like Herman Kahn who assert that some nuclear wars would be survivable, but overlooks arguments from what Glaser (1989) terms the *punitive retaliation* school of thinkers who suggest that the costs of a nuclear war are so high that the difference between 10 nuclear attacks and 100 or 1000 is negligible. If leaders' overriding concern is to retain

[19] We choose to present regression coefficients rather than predicted first-differences to improve the readability of the results. While predicted first-differences are a more substantively interesting quantity, there is too much estimation variability to make the predicted intervals meaningfully interpretable. Confidence bounds on some effects range from nearly 0 to upwards of a 80% increase in victory probability. The dataset is too small. Therefore, we focus simply on the rejection of the sharp null hypothesis of "no effect" rather than on estimating the precise impact of the independent variables on victory probabilities.

power, then the destabilizing impact of *any* nuclear conflict represents the ultimate cost. As Jervis (1979) argues, "almost no decision maker in the world's history would embark on a course of expansion while his cities were held hostage." A minor edge in superiority is insufficient to lower the costs of escalation.

What matters then is not a quantitative advantage in nuclear weapons but a qualitative advantage in nuclear posture - an assured second-strike. As Kroenig argues, the possibility of launching a successful counterforce strike against an adversary lowers the cost of nuclear "disaster." There is a greater incentive for a counterforce-capable state to push a competition in risk-taking forward since the event of a nuclear exchange is manageable. But if we assume that leaders are sufficiently risk-averse that they fear any reasonably-sized nuclear attack on their population, then a counterforce capability must be able to neutralize the near entirety of an adversary's arsenal in order to justify more aggressive crisis bargaining.

In short, a state that does not have a secure arsenal will likely have less resolve in a crisis since its relative cost of escalation is much higher. But if the adversary has a guaranteed ability to retaliate, typically via a highly-mobile nuclear force, then there is no meaningful difference in resolve between the two nuclear actors that can be explained by their respective nuclear capabilities. Empirically, this is the more consistent trend - a survivable deterrent increases the likelihood of being victorious in a crisis.

7 Conclusion

What is a reasonable number of nuclear weapons for a state to maintain is a crucial and highly politicized question. Kroenig contributes to this debate by arguing for the advantages of a larger nuclear arsenal, finding that nuclear superior states – those with a greater number of nuclear weapons – prevail more often than nuclear inferior states in international crises. However, upon testing this argument empirically, we are unable to conclude that nuclear superiority has an effect on a state's success during a crisis. Once we employ more accurate standard error corrections for such a small sample size, the effect of a larger arsenal becomes statistically insignificant at a reasonable level. Additionally, we find that the author's results are driven by the decisions made about how to code and model the data. We conduct a thorough review of the crises potentially involving two nuclear powers and propose alterations – guided by historical experts and the theoretical quantity of interest – to those original coding decisions. After exploring variations in how these historical events may be interpreted and different specifications of the model, the most robust finding is that survivability rather than superiority results in greater average success in crisis bargaining. Despite the consistency of our finding among the given observations, because the world has experienced very few crises among nuclear powers we argue for extreme caution in drawing any conclusions about the value of nuclear weapons. Regardless, our results certainly cast doubt on the desirability of a large nuclear arsenal, especially given the costs of maintaining nuclear superiority. Decisions about the optimal amount of nuclear weapons should be based on the value that each additional nuclear weapon provides relative to its costs. Our findings suggest that there are substantially diminishing marginal returns to nuclear development past the level of assured second-strike.

References

Aftergood, S. & Kristensen, H. (2007), 'Israel: Nuclear weapons', Federation of American Scientists.

Albright, D. (2000), 'India's and Pakistan's fissile material and nuclear weapons inventories, end of 1999', Institute for Science and International Security.

Angrist, J. & Lavy, V. (2009), 'The effects of high stakes high school achievement awards: Evidence from a randomized trial', *The American Economic Review* **99**(4), 1384–1414.

Appleman, R. E. (1989), *Disaster in Korea: The Chinese Confront MacArthur*, Texas A&M University Press.

Arellano, M. (1987), 'Computing robust standard errors for within-groups estimators', *Oxford Bulletin of Economics and Statistics* **49**(4), 431–434.

Bell, R. M. & McCaffrey, D. F. (2002), 'Bias reduction in standard errors for linear regression with multi-stage samples', *Survey Methodology* **28**(2), 169–182.

Cameron, A. C., Gelbach, J. B. & Miller, D. L. (2008), 'Bootstrap-based improvements for inference with clustered errors', *The Review of Economics and Statistics* **90**(3), 414–427.

Cirincione, J., Wolfsthal, J. B. & Rajkumar, M. (2002), *Deadly Arsenals*, Carnegie Endowment for International Peace.

Clarke, S. (1968), *The Congo Mercenary: A History and Analysis*, South African Institute of International Affairs.

Gartzke, E. A. & Gleditsch, K. S. (2008), 'The ties that bias: Specifying and operationalizing components of dyadic dependence in international conflict', Working Paper.

Glaser, C. (1989), Why do strategists disagree about the requirements of strategic nuclear deterrence, *in* L. Eden & S. E. Miller, eds, 'Nuclear Arguments: Understanding the Strategic Nuclear Arms and Arms Control Debates', Cornell, pp. 109–171.

Gleijeses, P. (1994), ‘Flee! the white giants are coming!: The United States, the mercenaries, and the Congo, 196465’, *Diplomatic History* **18**(2), 207–237.

Hausman, J. (2001), ‘Mismeasured variables in econometric analysis: Problems from the right and problems from the left’, *The Journal of Economic Perspectives* **15**(4), 57–67.

Hoff, P. D. & Ward, M. D. (2004), ‘Modeling dependencies in international relations networks’, *Political Analysis* **12**(2), 160–175.

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *in* ‘Proceedings of the Fifth Berkley Symposium in Mathematical Statistics’, Vol. 1, University of California Press, pp. 221–233.

Imbens, G. W. & Kolesar, M. (2012), ‘Robust standard errors in small samples: Some practical advice’, National Bureau of Economic Research Working Paper.

Jervis, R. (1979), ‘Why nuclear superiority doesn’t matter’, *Political Science Quarterly* **94**(4), 617–633.

King, G., Honaker, J., Joseph, A. & Scheve, K. (2001), ‘Analyzing incomplete political science data: An alternative algorithm for multiple imputation’, *American Political Science Review* **95**(1), 49–69.

King, G. & Zeng, L. (2006), ‘The dangers of extreme counterfactuals’, *Political Analysis* **14**(2), 131–159.

King, G. & Zeng, L. (2007), ‘When can history be our guide? the pitfalls of counterfactual inference’, *International Studies Quarterly* **51**(1), 183–210.

Kroenig, M. (2013*a*), ‘Debating the benefits nuclear superiority for crisis bargaining, Part i’, The Duck of Minerva.
URL: *http://www.whiteoliphaunt.com/duckofminerva/2013/03/debating-the-benefits-of-nuclear-superiority-part-iii.html*

Kroenig, M. (2013*b*), ‘Nuclear superiority and the balance of resolve: Explaining nuclear crisis outcomes’, *International Organization* **67**(1), 141–171.

MacKinnon, J. G. & White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**, 305–325.

McCaffrey, D. F., Bell, R. M. & Botts, C. H. (2001), Generalizations of biased reduced linearization, *in* 'Proceedings of the Annual Meeting of the American Statistical Association'.

National Security Archive (2006), 'Record of National Security Council meeting held on May 24, 1967 at 12 noon – discussion of Middle East crisis'.
URL: *http://www.gwu.edu/ nsarchiv/israel/documents/misc/02-01.htm*

Norris, R. & Kristensen, H. (2009), 'Nuclear notebook: Worldwide deployments of nuclear weapons, 2009', *Bulletin of the Atomic Scientists* **65**(86).

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley.

Russell, R. (2000), 'The 1996 Taiwan Strait crisis: The United States and China at the precipice of war?', Pew Case in International Affairs 231, Georgetown University Institute for the Study of Diplomacy.

Ryabushkin, D. (2007), Origins and consequences of the Soviet-Chinese border conflict of 1969, *in* A. Iwashita, ed., 'Eager Eyes Fixed on Eurasia', Sapporo.

Sechser, T. & Fuhrmann, M. (2013*a*), 'Debating the benefits of nuclear superiority for crisis bargaining, Part iii', The Duck of Minerva.
URL: *http://www.whiteoliphaunt.com/duckofminerva/2013/03/debating-the-benefits-of-nuclear-superiority-part-iii.html*

Sechser, T. S. & Fuhrmann, M. (2013*b*), 'Crisis bargaining and nuclear blackmail', *International Organization* **67**(1), 173–195.

Tellis, A. (2001), 'India's emerging nuclear posture', RAND Research Brief.
URL: *http://www.rand.org/pubs/research_briefs/RB63.html*

Treier, S. & Jackman, S. (2008), 'Democracy as a latent variable', *American Journal of Political Science* **52**(1), 201–217.

White, H. (1980), ‘A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity’, *Econometrica* **48**, 817–838.

White, H. (1984), *Asymptotic Theory for Econometricians*, Academic Press.

Windrem, R. & Kupperman, T. (2000), ‘Pakistan’s nukes outstrip India’s, officials say’, MSNBC.com.

Yang, K. (2000), ‘The Sino-Soviet border clash of 1969’, *Cold War History* **1**(1), 21–52.

Appendix A: Degrees of Freedom Corrections for Cluster-Robust Standard Errors

In the simple linear regression case, the variance-covariance matrix of the estimated regression coefficients, $V(\hat{\beta})$ is written as

$$V(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Sigma\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

If we assume homoskedasticity, $\Sigma = \sigma^2\boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix. However, if errors are heteroskedastic, Σ becomes a diagonal matrix with $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_n^2)$ on the diagonal. Huber (1967) and White (1980) show that in the presence of heteroskedasticity,

$$V(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^{n}[\boldsymbol{X}_i' e_i e_i' \boldsymbol{X}_i](\boldsymbol{X}'\boldsymbol{X})^{-1}$$

is a consistent estimator of the variance-covariance matrix where e_i is the residual for observation i. White (1984) and Arellano (1987) extend the heteroskedasticity-consistent estimator to the case where observations may be correlated within clusters of observations but independent across clusters. The cluster-robust standard error estimator for linear regression is

$$V(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^{M}[\boldsymbol{X}_i' \mathbf{e}_i \mathbf{e}_i' \boldsymbol{X}_i](\boldsymbol{X}'\boldsymbol{X})^{-1}$$

where M is the number of clusters, $\mathbf{e}_i$ is the vector of residuals for the observations in cluster i and X_i is a c_i by k matrix of covariate values, c_i denoting the number of observations in cluster i and k the number of covariates.

With the *HC1* multiplicative bias correction, the variance-covariance matrix is

$$V_{HC1}(\hat{\beta}) = \left(\frac{M}{M-1}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^{M}[\boldsymbol{X}_i' \mathbf{e}_i \mathbf{e}_i' \boldsymbol{X}_i](\boldsymbol{X}'\boldsymbol{X})^{-1}$$

The *HC2*-corrected CRSE estimator as defined in Bell & McCaffrey (2002) and mentioned in

Imbens & Kolesar (2012) is

$$V_{HC2}(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^{M}[\boldsymbol{X}_i'(\boldsymbol{I}_{c_i} - \mathbf{H}_{ii})^{-\frac{1}{2}}\mathbf{e}_i\mathbf{e}_i'\left((\boldsymbol{I}_{c_i} - \mathbf{H}_{ii})^{-\frac{1}{2}}\right)'\boldsymbol{X}_i](\boldsymbol{X}'\boldsymbol{X})^{-1}$$

where $\mathbf{H}_{ii} = X_i(X'X)^{-1}X_i'$ and $\boldsymbol{I}_{c_i}$ is a c_i x c_i identity matrix.

While the above CRSE estimators cover linear models, it is not difficult to extend them to non-linear models like the probit model used in Kroenig (2013*b*). For generalized linear models estimated with iteratively reweighted least squares (IRLS) and Fisher scoring, we can rewrite the covariance estimators and include the working weights and working residuals from the last stage of GLM estimation.[20] McCaffrey et al. (2001) explain a simple tweak to obtain cluster-robust covariance matrices for GLM.

1. Let $\boldsymbol{X}^* = (\mathbf{W})^{\frac{1}{2}}\boldsymbol{X}$ and $\mathbf{e}^* = (\mathbf{W})^{\frac{1}{2}}\mathbf{e}_w$ where $\mathbf{e}_w$ is the vector of working residuals, and $\mathbf{W}$ is a square diagonal matrix with the working weights on the diagonal.

2. The *HC2*-corrected variance-covariance matrix for the non-linear model coefficients can be estimated by replacing all instances of $\boldsymbol{X}$ and $\mathbf{e}$ in the above estimator formulas with $\boldsymbol{X}^*$ and $\mathbf{e}^*$.

[20] The `glm()` function in the `R` statistical software package estimates models using IRLS/Fisher scoring and returns vectors of working weights and residuals by default