by Matthew Fuhrmann, Texas A&M University, Matthew Kroenig, Georgetown University, and Todd S. Sechser, University of Virginia

In “What We Talk About When We Talk About Nuclear Weapons: A Review Essay,” Francis J. Gavin reviews and evaluates the recent quantitative turn in nuclear security studies, singling out for special attention our recent articles in *International Organization*, “Nuclear Superiority and the Balance of Resolve: Explaining Nuclear Crisis Outcomes” and “Crisis Bargaining and Nuclear Blackmail.”² The scope of Gavin’s critique extends well beyond our articles, however, encompassing the broader scholarly literature employing quantitative analysis to understand questions of nuclear security. “Statistical analysis,” he writes, “does not strike me as the best method for understanding complex, interactive political decision-making about issues of life and death.” Instead, Gavin argues for a research agenda dominated by qualitative methods and analysis of archival evidence to better understand the effects of nuclear weapons on international politics.

Gavin’s critique arrives at a pivotal time for the field of nuclear security. Scholars are increasingly turning to quantitative methods to answer questions about the political dynamics of nuclear weapons.³ Given the importance of the subject matter, it is crucial that nuclear security scholars carefully scrutinize their research methods. If Gavin is correct, much intellectual firepower is being wasted – with potentially serious consequences.

¹ The authors’ names are listed alphabetically; equal authorship is implied. The authors would like to thank Jason Gusdorf for helpful research assistance.


The issues Gavin raises are not unique to the field of nuclear security. The value of statistical analysis has been a topic of heated discussion for years in political science, sociology, anthropology, and other disciplines traditionally dominated by qualitative approaches. Beyond academia, quantitative analysis is being used today for corporate employee evaluation, the targeting of digital advertisements, political campaign management, online product recommendations, and many other areas of our lives. These debates have even seeped into professional sports: baseball and other sports have undergone an upheaval in recent years as quantitative methods have revolutionized the way teams think about strategy, scouting, and compensation.4 Gavin’s criticisms of these methods therefore have far-reaching implications, and deserve to be taken seriously.

We appreciate Gavin’s engagement with our work, and it is a privilege for us to participate in this important roundtable. As regular readers of this roundtable series already know, Gavin’s work on nuclear proliferation has helped reshape our understanding of the ways in which nuclear weapons shape world politics.5 In our view, however, Gavin’s critiques are badly misguided. In this essay, we defend quantitative analysis as an important tool for illuminating the complex effects of nuclear weapons. Contrary to Gavin’s claims, statistical analysis has several useful attributes for studying nuclear security. Instead of revealing the flaws of quantitative analysis, Gavin’s critiques suggest a misunderstanding of the nature and purpose of quantitative research, particularly as applied to nuclear security studies. Moreover, the alternative Gavin proposes would not solve the problems he highlights, and in some cases would exacerbate them. While the authors of this essay disagree among ourselves about many important substantive issues, we are united in the belief that statistical analysis has an important role to play in the field of nuclear security studies.

The rest of the essay will continue in four parts. First, we briefly describe our approach to studying nuclear issues, and contrast it with Gavin’s method. Second, we explain the unique advantages offered by statistical analysis. Next, we discuss the limitations of this approach and the complementary strengths of alternative methods. Finally, we offer concluding remarks.

---


We agree with Gavin on an important point: the study of nuclear weapons in world politics is important. Indeed, there are few issues of greater policy significance than the causes and effects of nuclear proliferation. All of the participants in this roundtable, therefore, want to better understand how nuclear weapons influence deterrence and coercion, and whether nuclear superiority provides countries with advantages in crisis bargaining. Yet we approach these issues from fundamentally different methodological perspectives.

Gavin’s proposed approach to studying these issues is straightforward: he suggests that we first “identify the most important example where these issues are engaged, look at the primary documents,” and then “reconstruct the thoughts, decisions, interactions between, and reconsiderations of top decision-makers” in order to determine whether nuclear weapons played a role in the crisis. In other words, Gavin argues that the best way to understand the political effects of nuclear weapons is to probe a single case (or a small number of cases) deeply. Specifically, he points to the 1958–1962 superpower standoff over Berlin and Cuba as the “most important and representative case” for studying nuclear crisis behavior, and returns to this episode repeatedly throughout the essay to support his arguments about nuclear deterrence and compellence.

Our respective articles in International Organization adopt a considerably different approach. Whereas Gavin focuses his attention on a single episode, our articles compare many episodes: Kroenig’s study evaluates 52 nuclear crisis participants, and Sechser and Fuhrmann analyze 210 coercive threats. For each study, we devise quantitative indicators for several key concepts—including nuclear capabilities, crisis actions, and several other factors. We then estimate statistical models to determine whether nuclear capabilities are reliably correlated with crisis outcomes, while controlling for other variables that could influence both a state’s nuclear status and whether it prevails in a crisis.

To illuminate the differences between our approaches, consider how we might approach an important question that is unrelated to nuclear security: does smoking cause cancer? While smoking and nuclear crises are vastly different phenomena, they share important similarities from a research standpoint. Both cigarettes and nuclear weapons have potentially significant effects on large numbers of people, but the ways in which they operate are not always visible to the naked eye. Further, conducting direct experiments to assess their effects is infeasible, whether due to ethical or practical reasons. Both questions therefore require us to think carefully about how to distill causal effects from imperfect data.

7 Ibid., p. 6.
8 This case is also a central subject of Gavin’s recent book, Nuclear Statecraft.
Using the “most important example” approach discussed by Gavin, the first step would be to locate the most important individual example of smoking and evaluate this person’s life. One problem with this approach emerges immediately: it is not clear who the most important smoker might be – perhaps an individual with a high public profile, one with a great deal of wealth, or someone who smoked excessively. As we will discuss below, identifying the most important nuclear crisis is likewise fraught with difficulty. Assuming we could identify this “most important” case, we might then ask, for example, how much this person smoked each day, what diseases or health problems he acquired, his opinion about how smoking affected his life and health, and at what age he died. We would also scour medical records and the results of medical tests for information about this person’s behavior and health over his lifetime. What did his doctors say about the effects of his smoking? Did they believe that smoking put him at a higher risk of getting cancer? What did this person himself believe? Do we have transcripts in which this person discusses his views on the health effects of his habit? With this trove of information, we could then formulate an understanding about smoking and cancer based on the written and verbal record of this person’s health.

By contrast, our approach would begin by collecting data on a large number of individuals, including smokers and nonsmokers as well as people with and without cancer. Across this large group of subjects, we would analyze whether there was a strong correlation between the “independent variable” (smoking) and the “dependent variable” (cancer). Are there patterns between smoking and cancer that cannot be explained by chance alone? If we found that smoking was statistically associated with a higher incidence of cancer – after controlling for other factors that affect whether one smokes and whether they contract cancer, such as family history, occupation, and other behaviors – we would then conclude that smoking is correlated with, and may therefore be a cause of, cancer.

Which approach described above is more useful for understanding the causes and effects of smoking? What about the causes and effects of nuclear proliferation? Gavin’s method is not without merit, but it suffers from major drawbacks that impede one from making reliable inferences about how nuclear weapons affect world politics. Our approach has important limitations as well, but we believe that it is a more powerful tool for providing answers to the questions we raise in our International Organization articles.

Using statistics to study nuclear security offers many advantages. The central goal in any social scientific study is inference: to observe a small slice of the world in order to learn how it works. Specifically, both of our studies in International Organization endeavor toward causal inference, aiming to learn about the causal effects of nuclear weapons in crisis situations. How do nuclear weapons shape the way crises play out?

---

9 We do not know for certain how Gavin would approach this question, of course, since he does not address the relationship between smoking and cancer in his essay. Our goal here is simply to discuss how his approach to studying nuclear security would apply in another context.

10 King, Keohane, and Verba distinguish between descriptive inference, which uses observable information to learn about unobserved facts (such as inferring one’s favorite sports team from one’s city of residence), and causal inference, in which the goal is to ascertain causal effects. See Gary King, Robert O.
Are nuclear states more successful when employing deterrence and coercion? Are states with a nuclear advantage more likely to achieve their basic political goals in a crisis?

Which method offers clearer insights into the questions driving our studies? Below we describe four key advantages of employing a quantitative approach to evaluating the political effects of nuclear weapons.

First, quantitative approaches allow us to compare large numbers of observations. Undoubtedly, Gavin’s approach to studying nuclear weapons is the better method for learning about the events of the 1958–1962 period. But does this single episode tell us all we need to know about nuclear weapons? Our International Organization articles ask broader questions about how nuclear weapons have impacted world politics in the half-century before and after this period – and how they might do so in the future. Claiming that the Berlin/Cuba episode is sufficient for answering these questions presumes that the decades before and since the Cuban missile crisis have little to teach about the consequences of nuclear weapons – a claim we find specious at best.

The main limitation of Gavin’s approach to assessing the effects of nuclear weapons is that it provides us with few, if any, generalizable inferences – the central aim of social science. In other words, his approach tells us much about the Cuba and Berlin crises, and rather little about nuclear weapons more broadly. Did the patterns that were evident from 1958–1962 continue throughout the Cold War, or were they anomalous? Should we expect that these patterns will hold into the future? Gavin’s approach offers no answer because it unduly limits its scope to a single historical episode, rather than examining how that episode fits into broader trends.

The questions we ask in our articles require a more comprehensive approach to data collection. By collecting information about dozens (or hundreds) of cases rather than just one or two, we can gain insights into whether the patterns we observe in any individual case are representative of broader trends. The implicit question in our research is always ‘what would have happened if conditions had been different?’ Of course, it is impossible to answer this counterfactual with certainty since history happens only once, and we cannot repeat the ‘experiment’ in a laboratory. But that does not mean we should shrug our shoulders and abandon the enterprise.

Instead, we can gain insight by looking at cases in which conditions were, in fact, different. To illustrate, let’s return to the smoking example above. Studying a single smoker in depth might give us an accurate and textured understanding of the role of smoking in this

---


11 Indeed, Gavin admits as much when he concedes that “Even if a deep immersion in documents produces historical accuracy, it often comes at the cost of the generalizations and policy insights about nuclear dynamics we all crave.” See Gavin, “What We Talk About When We Talk About Nuclear Weapons,” 21.
person’s life, but it would be a poor way to learn about the broader health effects of smoking, because we could not make an informed guess about what would have happened had he not smoked. Our approach described earlier, in contrast, allows us to generalize about the effects of smoking on health. For precisely this reason, large-scale quantitative analysis is the primary method by which medical researchers have tackled the health effects of tobacco smoke. To be sure, some of the data in our hypothetical study would surely be inaccurate, and we would know comparatively little about the lives of each individual subject. But the loss in individual case knowledge would be more than compensated by the increase in information about the variables we hope to study.

So it is with nuclear weapons. To understand how nuclear weapons impact international crises, we must examine crises in which nuclear ‘conditions’ were different. For Kroenig, this means comparing the fortunes of crisis participants that enjoyed nuclear superiority to those that did not. For Sechser and Fuhrmann, it means comparing the effectiveness of coercive threats made by nuclear states to those made by nonnuclear states. By making these comparisons, we can begin to engage in informed and evidence-based speculation about how nuclear weapons change (or do not change) crisis dynamics. Indeed, the statistical models we employ require this comparison – they will return no results if all of our cases look the same.

Gavin argues that the Berlin/Cuba episode is sufficient for understanding the dynamics of nuclear weapons because it is the “most important and representative” case of nuclear deterrence and coercion. There are two distinct (and contradictory) claims here: that the case is the most important crisis episode for studying nuclear weapons, and that it is representative of the broader universe of such episodes. With respect to the first claim, Gavin offers no criteria for evaluating what an “important” case might be. What makes a case important – its profile among the general public? Its consequences? The availability of information about it? The countries involved? Moreover, for whom must the case be important? Gavin may view the 1958–1962 case as critical for understanding nuclear dynamics, but it is by no means clear that policymakers today look to this example for guidance about dealing with Iran or North Korea. This is not to say that we disagree with Gavin’s assessment – undoubtedly the 1958–1962 episode is important in many respects. But importance, like beauty, is in the eye of the beholder.

The second claim is equally dubious: that the 1958–1962 episode is somehow representative of the ways in which nuclear weapons typically shape international politics. Without first examining other cases, Gavin simply has no grounds on which to base this claim. Moreover, there is tension between this claim and his previous assertion that the case is important: one key reason the Cuba/Berlin episode is often seen as important is because it was not like other Cold War crises: nuclear weapons were brandished more explicitly, and stoked more public anxiety about nuclear war, than any other crisis before or since. In the broader universe of crises, this episode actually may be quite anomalous. If

---

12 Ibid., 6.
so, then studying it to the exclusion of other cases would yield misleading conclusions about the role of nuclear weapons in world politics.

A key advantage of quantitative methods is that the researcher need not make questionable judgments about which cases are more or less important: unless explicitly instructed otherwise, statistical models assign equal weight to each case. Likewise, statistical models provide ways to identify – and exclude – anomalous cases that deviate markedly from dominant trends. Indeed, a quantitative analysis can be a useful precursor to the selection of individual cases for in-depth analysis, precisely because it allows us to locate cases that either represent or deviate from the overall pattern. These selections, however, are based on careful comparisons with other cases, not opaque judgments.

A second advantage is that quantitative analyses provide greater transparency about methods, judgments, and conclusions. One of Gavin’s central critiques is that various cases in our quantitative analyses have been miscoded. In other words, he argues, we have mismeasured important factors.13 This criticism – irrespective of its validity14 – is possible only because our coding decisions are unambiguous and easily ascertained from our datasets. Moreover, each of our studies sets forth clear rules for how each variable in our datasets was coded. This does not mean that our coding decisions are all correct and beyond dispute, but it does mean that they are clearly stated for outside scholars to evaluate. This degree of transparency is a key strength of quantitative research. Because each case in a quantitative analysis necessarily must be clearly coded,15 there is no ambiguity about how the researcher has classified each case. If other researchers believe a case should be coded differently, they can make that change and rerun the analysis.

By extension, quantitative research designs permit scholars to easily evaluate how much a study’s findings depend on individual coding decisions. Simply noting a few coding errors or differences of interpretation in a large quantitative dataset is of little consequence unless one can demonstrate that those differences are responsible for generating incorrect inferences. In a quantitative study, this typically amounts to recoding disputed cases and repeating the core statistical models to determine whether the results change substantially.16 Not only are the original coding decisions laid bare, but it is also


14 In our individual responses to Gavin’s essay, we address the particulars of these coding disagreements.

15 In the quantitative models we use, cases are dropped from the statistical analysis if even one variable is not coded.

16 For example of a case where such a reassessment yields substantially different findings, see Alexander B. Downes and Todd S. Sechser, "The Illusion of Democratic Credibility," International Organization, Vol. 66, No. 3 (2012): 457-89.
straightforward to determine whether the study’s inferences depend on them. This high level of transparency – and the external quality-control it enables – is one of the most attractive features of quantitative research designs. Transparency is useful not because it produces scholarly consensus, but because it allows opposing sides to identify the precise nature and implications of their disagreements.

Consider, for example, the 1990 exchange in World Politics between Paul Huth and Bruce Russett on one hand, and Richard Ned Lebow and Janice Gross Stein on the other.17 Gavin highlights the similarities between this debate and the present exchange, separated by almost twenty-five years, as evidence that quantitative analysis has made little progress in understanding nuclear issues. We see the issue differently. Both debates, in fact, illustrate a key strength of quantitative analysis: the ability to assess the importance of individual coding decisions. In the World Politics debate, Lebow and Stein objected that Huth and Russett had improperly coded many cases in their deterrence dataset, much as Gavin has disputed some of our classifications But Huth and Russett responded by noting that “even if Lebow and Stein’s recodings of our cases are accepted, the statistical and substantive findings of our past research remain fundamentally unchanged.”18 Similarly, as we report in our articles, our central findings do not change even if we accept Gavin’s arguments. In a quantitative study, simply showing that certain coding decisions can be contested is insufficient: one must also demonstrate that the core results depend on those decisions. While Gavin is correct to argue that coding cases is a tricky exercise, quantitative approaches allow us to evaluate the substantive importance of questionable coding decisions.

Qualitative research, by contrast, is not always so amenable to external oversight. Whereas quantitative models demand clear coding decisions, qualitative research designs can be much more forgiving of ambiguous classifications. Gavin’s critique of our coding decisions illustrates this problem: while he criticizes the way we have coded particular cases in our datasets, he offers no clear alternative coding scheme. He raises questions about our coding decisions, but then declines to answer them. This ambiguity allows him to have his cake and eat it too: he can criticize our classifications without being liable for his own.

Uncertainty, of course, is inherent to any scientific enterprise, and quantification is sometimes criticized for presenting a false illusion of certainty. To be clear, quantitative research cannot create certainty where the evidence is ambiguous. Just because a case is coded a certain way does not mean that the broader scholarly community (or even the researcher) has reached a consensus about that case. Likewise, the problem of ambiguity is not inherent to qualitative research: nothing intrinsic to historical research precludes

---


18 Thanks to Erik Gartzke for pointing this out. See Huth and Russett, ”Testing Deterrence Theory,” 468.
scholars from laying their assumptions bare. But by compelling scholars to take a clear initial position on coding cases, the process of quantification allows scholars to debate each decision and evaluate whether potentially questionable choices are decisive in generating a study’s core results. This transparency is central to peer evaluation and, ultimately, scientific advancement.

A third advantage of statistical analysis is that it is designed to cope with probabilistic events. In the physical world, causal relationships are often deterministic: a certain amount of force imparted to an object will cause that object to move a certain distance. So long as conditions are kept constant, this result will obtain again and again, no matter how many times the experiment is repeated. In the social world, however, we are not blessed with such ironclad reliability. No two individual people are exactly identical, and even in carefully controlled environments it is rare to find a “force” that begets exactly the same effect on all people with perfect regularity. The causal relationships we observe are not deterministic – they are probabilistic, occurring with imperfect regularity.19

The ‘force’ of interest to us in our articles is, broadly, the possession of nuclear weapons. When this force is applied to crisis bargaining situations, what happens? Implicit in this question, however, is a question about probability: when nuclear weapons are inserted into a crisis bargaining situation, what is the likelihood of a particular outcome? Kroenig’s study, for example, asks: in a nuclear crisis, what is the likelihood that the nuclear-superior side will achieve its basic goals? Likewise, Sechser and Fuhrmann seek to discover the likelihood that a coercive demand made by a nuclear-armed state will be met.

The central difficulty with posing our research questions in this way is that we cannot actually see the thing we care about: probability is inherently unobservable. We cannot examine a crisis and directly observe the probability of one side capitulating; we can only observe whether it actually capitulated.20 How, then, can we begin to answer our original research question?

Quantitative research is designed for precisely this sort of situation. If we cannot directly

19 King, Keohane, and Verba note that there are two possibilities here: either the social world truly is probabilistic, or it is actually deterministic and simply appears probabilistic to us because our explanations are imperfect. These two possible worlds, they note, are indistinguishable to our eyes – in other words, probabilistic theories are equally useful for describing phenomena in either one. See King, Keohane, and Verba, Designing Social Inquiry, pp. 59-60.

20 Participants in crises sometimes offer their own probability assessments. President John F. Kennedy famously estimated the odds of nuclear war during the Cuban missile crisis as "somewhere between one in three and even." See Theodore C. Sorensen, Kennedy (New York: Harper & Row, 1965), 795. Yet there is no reason to believe that participants, operating in the heat of the moment and with limited information, would have any better access to underlying probabilities than a disinterested analyst conducting an assessment after the fact. Further, participants often disagree sharply: Kennedy’s National Security Adviser, McGeorge Bundy, estimated the risk of war to be just one in 100 (though “still much too large for comfort”). See McGeorge Bundy, Danger and Survival: Choices About the Bomb in the First Fifty Years (New York: Random House, 1988), 461.
observe whether we are holding a loaded six-sided die, for example, we can throw it many times, observe the result, and infer the underlying probability from the results. Throwing the die just one time would tell us little, since all six numbers are theoretically possible even if the die were loaded. Only after observing the pattern of results across many events can we determine the underlying probabilities of each number turning up.

The single-case approach Gavin proposes cannot cope with probabilistic events as effectively. Knowing that one smoker happened to die of cancer does not tell us much about the broader health effects of tobacco. Based on this single data point, we might conclude that smoking leads to cancer 100 percent of the time. Yet we know this to be false: there are heavy smokers who remain cancer-free, just as there are nonsmokers who still get cancer. The true relationship between smoking and cancer emerges only after looking at a large number of cases. Similarly, even if we determine that nuclear weapons appeared to “matter” from 1958-1962, we cannot safely infer from this observation that nuclear weapons influence crisis outcomes in general. Any relationships observed during this particular period could have been due to any number of chance events that might be unlikely to recur. Studying just one episode allows us to say much about that episode but little about the underlying relationships.

Fourth, statistical analysis allows researchers to uncover causal relationships in social phenomena even if the participants themselves do not record, record accurately, or understand these relationships. Gavin’s approach, in contrast, requires finding primary source documents and learning what participants themselves believed to be the relevant causal factors at play. His essay conveys an exceptionally narrow conception of how one should gather knowledge about the effect of nuclear weapons on international politics. Gavin believes that if one wants to “really understand” the effect of nuclear weapons on international politics, archival research is “the only way to get real insight.” While we agree that studying primary documents has great value, we believe that there are many other ways to generate useful knowledge, and that a narrow focus on primary documents can often lead a scholar astray.

A focus on primary documents alone has several problems. First, the historical record is often incomplete, but the absence of evidence is not evidence of absence. In any historical nuclear crisis, a leader might have made an important statement to his advisers about how nuclear weapons fundamentally affected a crucial decision, but the statement might have never been recorded or could have been lost in the sands of time. Without that crucial piece of information, one might conclude, incorrectly, that nuclear weapons were not part of the calculation. Second, and related, something might not appear in the historical record because it was taken for granted. Researchers might then conclude that this factor was not salient, when in fact it might have been so important that it was well understood by everyone in the room and did not need to be mentioned. For example, nuclear weapons


22 Ibid., 16. Italics in original.
may have been so irrelevant (or central) to a particular crisis that leaders didn’t even see the need to raise the issue. Third, individuals sometimes intend to mislead. Leaders may make inaccurate or incomplete statements – even in private – in order to influence the outcome of an internal debate, to improve domestic political fortunes, to shape how they will be viewed by posterity, or a variety of other reasons. Fourth, something might not appear in historical documents because participants themselves were unaware of how important it was. A case study of a smoker in the 1920s might not turn up any evidence from the smoker's medical records that smoking was damaging to his health. But, it was damaging nonetheless. Similarly, nuclear weapons may have had an influence on crisis dynamics even if leaders themselves didn’t fully appreciate it.

Statistical analysis, on the other hand, does not depend on the participants themselves to understand and record accurately the causal forces at work. Rather, the researcher can independently identify and measure the variables of interest and search for correlations in large samples of data.

In short, factors may matter regardless of whether the participants record them, record them accurately, or even understand their existence. Statistical analysis can reveal these connections. The documentary record might not.

Statistical analysis can help us understand world politics, but it is not without limitations. Like any method, large-N analysis has some potential pitfalls with which scholars must grapple. In this section we review some of the key limitations of this approach and describe how we cope with them in our studies.

Our goal as social scientists is to understand causal relationships between variables of interest. However, it is difficult to determine causality in studies that use observational data, like ours. Why is this the case? When we find a strong correlation between variables A and B in our observations of the real world, this does not necessarily mean that A “causes” B. If a researcher found, for example, a positive and statistically significant correlation between ice cream consumption and drowning, this would not necessarily imply that eating ice cream causes people to be more susceptible to drowning. Rather, it is more likely that a third variable – temperature – was responsible, since in warm weather people are more likely both to go swimming – which increases the risk of drowning – and eat ice cream.

There is a similar risk of drawing spurious relationships when studying the political effects of nuclear weapons. Imagine that one discovered a positive correlation between nuclear weapons and military conflict. This would not necessarily imply that nuclear weapons cause states to behave more belligerently. Nuclear proliferation occurs as a result of a strategic process: countries possess atomic bombs only after governments decide that it is in their interest to build them. The factors causing states to build the bomb – rather than the weapons themselves – might account for the observed relationship between bomb possession and conflict. In particular, states in threatening security environments may be more likely to seek atomic bombs to protect themselves. In this case, the dangerous
security environment is causing the state to both build nuclear weapons and become embroiled in a large number of disputes.\(^{23}\) What can be done to address this issue?

The best way to deal with it, from the perspective of causal inference, would be to create an experimental world in which no state possessed nuclear weapons, artificially manufacture dozens of crises, and then randomly assign nuclear weapons to some states. We could then compare whether states that received bombs (the treatment group) were more conflict-prone than countries that did not get weapons (the control group). If they were, we could reasonably conclude that nuclear weapons caused countries to become more belligerent, since a state’s nuclear status emerged by chance – not as a result of a strategic process. Of course, this approach is not only impossible, but also dangerous. We therefore must rely on other solutions to address barriers to inference that arise in observational studies.\(^{24}\)

The most straightforward way to reduce the risk that an observed relationship is spurious is to control for the factors that might be correlated with both the independent and dependent variable. Our studies account for several covariates – including conventional military power – that could affect crisis bargaining and the nuclear status of the crisis participants. However, in some cases, the relevant factors for which one needs to control may not be observable or measurable. Political scientists are increasingly turning to ‘high-tech’ tactics for addressing concerns about causality.\(^{25}\) Yet there are ‘low-tech’ solutions too that may be equally (and sometimes more) effective. Developing sound theory can go a long way in dealing with this issue. If an analyst has a logically consistent argument that can plausibly explain an observed correlation, she might have more liberty to make causal claims about that relationship. This would be especially true if she could use theory to rule out alternative explanations that could account for the observed pattern. In addition, supplementing statistical findings with case studies that unpack causal mechanisms can be an especially fruitful strategy.\(^{26}\)

\(^{23}\) See Bell and Miller, “Questioning the Effect of Nuclear Weapons on Conflict.”

\(^{24}\) Scholars may, however, conduct experiments in which they manipulate the treatment. For example, Daryl Press, Scott Sagan, and Benjamin Valentino recently conducted a survey experiment to assess why people oppose the first use of nuclear weapons. See Daryl Press, Scott D. Sagan, and Benjamin Valentino, “Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-Use of Nuclear Weapons,” *American Political Science Review*, Vol. 107, No. 1 (2013): 187-206. A natural drawback of this method, of course, is limited external validity; insights from the laboratory do not always translate to the real world.

\(^{25}\) For example, matching analysis is increasingly used in international relations research. Through this analysis, scholars construct samples in which the ‘treatment’ and ‘control’ groups are as similar as possible before estimating their statistical models. See, for example, Yonatan Lupu, “The Informative Power of Treaty Commitments: Using the Spatial Model to Address Selection Effects,” *American Journal of Political Science*, Vol. 57, No. 4 (2013), 912-925.

\(^{26}\) This is a strategy that we have pursued in our own work. See Kroenig, *Exporting the Bomb;* Fuhrmann, *Atomic Assistance;* and Todd S. Sechser, “Goliath’s Curse: Coercive Threats and Asymmetric Power,” *International Organization*, Vol. 64, No. 4 (2010), 627-660.
It is important to note that challenges related to causal inference are hardly unique to our methodological approach. Establishing causation in the social sciences is notoriously difficult, and it is a problem that plagues not just statistics, but many research methods. Gavin’s preferred method is particularly ill equipped for assessing causality. It is exceedingly difficult to know whether nuclear superiority caused the United States to win crises from 1958 to 1962 by following his approach. What we really want to know, if our goal is to make causal claims, is whether the outcomes of the crises during that period would have been different had the Soviets possessed strategic superiority instead. Of course, there is no way to rerun history, as we noted earlier in this essay. Yet one could get at this indirectly by comparing the 1958–1962 period to later crises in which the United States had nuclear inferiority but which were similar to the Berlin/Cuba cases in other respects. Gavin does not explicitly recommend this type of comparative analysis (although his discussion of the 1971 agreement on Berlin implicitly underscores the importance of controlled comparisons). If he had, however, our central point here would remain the same: conducting qualitative analysis does not exonerate scholars from wrestling with the issue of causality.

Readers might respond that Gavin’s approach is not designed to make causal claims, and that one’s ability to make such inferences using any method is exceedingly limited. At least one statement in Gavin’s essay implies that he accepts this view. As he writes, “For those that would complain that such indeterminacy undermines the idea of a political ‘science,’ I would respond, guilty as charged.” This is a perspective that we do not share, and neither do the vast majority of political scientists – including many who rely exclusively on qualitative methods.

A second limitation of statistical analysis has to do with measurement: it is sometimes difficult to quantify concepts of interest. Prestige, for example, is widely believed to be important in world politics. Indeed, according to the prevailing wisdom, one reason states seek nuclear weapons is to increase their standing in the international system. Yet it is tough to assign a numerical value to prestige. We can devise variables to proxy for this concept – for instance, one recent study codes status seeking behavior based on performance at the Olympics – but we cannot measure it directly. It is therefore difficult to test theories about prestige using statistical analysis. Measurement challenges likewise accompany other concepts in international relations: reputation, influence, soft power, etc.

Gavin zeroes in on this limitation when critiquing our research. He argues that seemingly straightforward phenomena, like who won a crisis, can be difficult to code. Consider the


Cuban Missile Crisis. Kroenig codes this case as a crisis victory for the United States and a defeat for the Soviet Union. Sechser and Fuhrmann similarly code the U.S. demand to withdraw the missiles as successful. However, Gavin argues that this coding is overly simplistic. Although the Soviets withdrew the nuclear weapons from Cuba in response to U.S. pressure, he suggests that both sides ultimately got what they wanted. The Soviets “won” because they preserved the Castro regime and forced the United States to remove Jupiter missiles from Turkey. Yet this nuance is lost in our quantitative analysis, according to Gavin.

We agree with Gavin that measurement issues can create challenges for scholars employing quantitative analysis. Yet qualitative research has to deal with these problems too. Let’s say, for the sake of illustration, that one wanted to design a qualitative study to assess whether nuclear superiority mattered in the Cuban Missile Crisis. Reliance on qualitative methods does not free an analyst from dealing with potentially thorny measurement issues. One would still need to know a minimum of two things: (1) which side won the Cuban missile crisis and (2) which side had nuclear superiority. It is impossible to conduct a proper test of the argument in the absence of this information. Any social scientific inquiry requires analysts to measure their key variables, whether they employ qualitative or quantitative analysis. The measurement problems that Gavin identifies undoubtedly exist to some degree in our studies, but they are by no means unique to our studies or to statistical analysis more generally.

Gavin also overstates the magnitude of these problems for our research when he implies that they impede our ability to glean any meaningful insights. To be sure, in some cases, who won a crisis may be ambiguous. But in many instances, which state was victorious is clearer than Gavin suggests. In 1979, for example, a crisis erupted after supporters of the Iranian revolution took 52 hostages at the American embassy in Tehran. The Iranian government refused to return the hostages even after the United States threatened it with military force, making this case an unambiguous failure for U.S. coercive diplomacy.

We also find Gavin’s interpretation of the outcome of the Cuban Missile Crisis difficult to sustain. We would not conclude that the Denver Broncos “won” Super Bowl XLVIII – a game they lost badly to the Seattle Seahawks – because their appearance in the game increased Broncos merchandise sales. The Broncos undoubtedly benefited in many ways from their Super Bowl appearance, but this hardly means that they won the game. Similarly, even if the Soviets got something out of the Cuban missile crisis, they failed to obtain their main objective – namely, keeping the missiles in Cuba. Our respective codings of this case reflect the most direct interpretation of the crisis’ outcome.

In some cases scholars who employ statistical analysis can cope with the measurement issues that Gavin identifies. If, for example, there is disagreement about the way a certain case is coded, scholars can simply recode it and see if the results change. To the extent that the findings are consistent regardless of how certain cases are coded, one can have greater confidence in the robustness of the results. As we explain above and below, our findings in these studies are robust to such changes.
A third potential concern is that statistical findings may be sensitive, especially when studying rare events. Students of international relations who employ statistical analysis face a key challenge: many of the events in which they are interested occur infrequently. This is particularly true in the area of nuclear proliferation, given that only 10 countries have built nuclear weapons. The rarity of nuclear proliferation does not mean that statistical analysis is useless for understanding how nuclear weapons affect world politics, as Gavin implies. It does mean, however, that scholars should exercise appropriate caution when using this tool. When dealing with rare events, there is always the possibility that statistical findings are driven by a small number of cases.

It is worth noting that the phenomena of interest in our studies – crises involving nuclear states – are not particularly rare. Kroenig’s study evaluates the performance of 52 nuclear crisis participants, and Sechser and Fuhrmann analyze 210 coercive threat episodes. Crisis participants achieved their basic goals in 35 percent of Kroenig’s cases, and in 30 percent of the cases evaluated by Sechser and Fuhrmann. Half of all crisis participants, by Kroenig’s definition, enjoyed nuclear superiority, while nuclear-armed challengers issued 20 percent of the coercive threats in Sechser and Fuhrmann’s study. In neither of our studies were the critical numbers “9, 2, and 0,” as Gavin suggests. In short, the rare-events problem is far less pernicious in our articles than Gavin implies.

Still, given that our samples are relatively small, the patterns we identified may be sensitive to minor changes in the statistical models. One way to deal with this issue is to conduct additional tests that are designed to evaluate the strength of a given finding, a solution similar to the one proposed above for addressing measurement problems. We employ this strategy in our articles – we rerun our models after recoding the dependent variables, modifying how we measure nuclear weapons and nuclear superiority, dropping potentially influential cases, and excluding particular control variables that could potentially bias our results. As we report in our respective studies, our main results survive these further tests. This does not guarantee that our findings are bulletproof, but it should inspire greater confidence that the patterns we identified are reasonably robust.

A fourth limitation is that statistical analysis is not designed to explain outliers. It is instead intended to identify average effects. Statistical analysis can tell us, on average, how an independent variable relates to a dependent variable. This is useful because it helps scholars to determine whether their theories are generalizable to a broad set of cases. Yet broad patterns may not apply to any single case. Most statistical models in political science produce outliers – cases that are not explained by the theory being tested. The 1999 Kargil War, for instance, is an outlier for the democratic peace theory. Although democracies are in general unlikely to fight militarized conflicts with one another, India and Pakistan nevertheless fought this particular war when they were both democratic states. But because most international relations theories are based on probabilistic logic (as opposed to deterministic logic), the presence of a few outliers does not necessarily disprove them.

It is often helpful to explicitly identify outliers, something that Kroenig (Table 1) and Sechter and Fuhrmann (Table 3) do in their articles. Scholars can then study these cases to refine their theories or, at the very least, identify the conditions under which their arguments hold. This type of ‘nested analysis’ – selecting cases for study based on results from a statistical model – is increasingly employed in international relations research to improve the explanatory power of theories. Identifying when outliers occur can be important for policy purposes. If we understand when nuclear powers were able to make successful threats in the past, for example, we can better understand whether Iran’s ability to blackmail its adversaries might change if it builds the bomb.

In an attempt to illustrate the folly of studying nuclear weapons with statistics, Gavin begins his article with an excerpt from a farcical story by Josh Freedman about a man breaking up with his girlfriend, Susan, as the result of “a series of quantitative calculations.” As the heartbreaker explains, “We can say we love each other all we want, but I just can’t trust it without the data.” The implication seems to be that certain things, like love and nuclear security, are simply too mysterious to understand using numbers.

Yet, in recent years, psychologists and sociologists interested in patterns of human courtship and mating have made major advances in their understanding of love and relationships through the employment of statistical analysis. Newly-available data from online dating sites (which themselves use quantitative data to power their pairing algorithms) have allowed scholars to test and refine their theories, and to develop new ones, in ways never before possible. For instance, the New York Times reports that scholars are now able to put a quantitative estimate on the value women put on a man’s professional success: “men reporting incomes in excess of $250,000 received 156 percent more e-mail messages than those with incomes below $50,000.” And this line of research has uncovered startling new findings, such as that white Americans, despite stated beliefs in racial equality, are especially unlikely to marry outside of their race. These are preferences that few people would admit to, or even be consciously aware of, if subjected to direct interviews by researchers, yet they become glaringly obvious in large patterns of data.

There are, then, similarities between love and nuclear security: statistical analysis can contribute to a fuller understanding of both. Like the study of human courtship, nuclear


security has undergone a renaissance in recent years, which has only been made possible by newly available datasets and a new generation of scholars applying statistical tools to address age-old questions. As we argue, these methods are advantageous for their ability to compare large numbers of observations, transparently reveal the researcher’s methods and judgments, deal with probabilistic phenomena, and uncover relationships about which participants themselves might not be cognizant.

Nevertheless, while we believe statistical analysis has important strengths, it cannot be the only instrument in our toolkit. Even with statistics, it is difficult to ascertain causation, key concepts may defy easy measurement, available data may be limited, and outliers will demand explanation. Yet, these drawbacks do not mean that statistical analysis should be abandoned altogether. Qualitative methods suffer from some of the same problems, and many others as well.

For decades, scholars have been employing the methods proposed by Gavin to study the role of nuclear weapons in international politics. Scholars have written dozens, if not hundreds, of studies assessing the consequences of nuclear weapons for deterrence and coercion. And yet, this approach has not lived up to its promise. Despite the availability of millions of pages of archival documents, our understanding of these issues remains inadequate. Indeed, Gavin’s own research (as he acknowledges) yields no conclusive answers about the coercive effects of nuclear weapons, and little useful policy advice, boiling down to the underwhelming finding that “U.S. nuclear superiority mattered. And, at some level, it also didn’t.”35 Faced with this situation, it would be unwise to respond by continuing to tackle questions of nuclear security in exactly the same way as before, hoping that the next trove of documents will settle the debate once and for all. Rather, we should consider the possibility that an incomplete methodological toolkit has been obscuring our view.

Matthew Fuhrmann is an Assistant Professor of Political Science at Texas A&M University. He has previously held research fellowships at Harvard University (2007-08) and the Council on Foreign Relations (2010-11). He is the author of Atomic Assistance: How “Atoms for Peace” Programs Cause Nuclear Insecurity (Cornell University Press, 2012) and the co-editor of The Nuclear Renaissance and International Security (Stanford University Press, 2013). His work has been published or conditionally accepted in peer reviewed journals such as American Journal of Political Science, British Journal of Political Science, International Organization, International Security, Journal of Conflict Resolution, Journal of Peace Research, and Journal of Politics.

Matthew Kroenig is an Associate Professor and International Relations Field Chair in the Department of Government at Georgetown University and a Nonresident Senior Fellow in

35 Philip Zelikow, H-Diplo Roundtable Reviews 15(1): 2013, 29. https://www.h-diplo.org/roundtables/PDF/Roundtable-XV-1.pdf. Gavin approvingly quotes this passage in his essay, leading us to wonder why he asks whether our work “resolve[s] the decades-long debate over these long-contested issues in nuclear dynamics” when he does not appear to hold his own to the same standard.

**Todd S. Sechser** is an Assistant Professor of Politics at the University of Virginia. His research interests include coercive diplomacy, reputations in international relations, the strategic effects of nuclear weapons, and the sources and consequences of military doctrine. His work has been published in peer-reviewed journals such as the *American Journal of Political Science, International Organization, International Studies Quarterly*, and *the Journal of Conflict Resolution*, and his commentary on international affairs has appeared in the *Wall Street Journal, Christian Science Monitor, Boston Globe*, and other outlets. He has held research fellowships at Stanford University and Harvard University, and from 2011-12 was a Stanton Nuclear Security Fellow at the Council on Foreign Relations. He is also a member of the Program on Strategic Stability Evaluation.
References


