Learning to Predict Proliferation

Nicholas L. Miller

Abstract How effective are states at assessing and predicting the nuclear intentions of foreign countries? Drawing on close to 200 US assessments of foreign countries' proliferation intentions between 1957 and 1966, this research note finds that close to 80 percent of testable US assessments were correct and that they shifted from highly inaccurate in the late 1950s to highly accurate in the 1960s. Based on quantitative and qualitative analysis, I conclude that learning from early failures led the intelligence community to achieve higher accuracy.

How effective are intelligence agencies at assessing and anticipating the nuclear intentions of foreign countries? More specifically, can they reliably detect and predict when other countries will seek to acquire nuclear weapons?

Correct assessments of this sort are a crucial precursor for successful nonproliferation efforts.[1] According to one recent study, the vast majority of nonproliferation deals are struck early, when a country is only considering a nuclear weapons program or has just begun one. If this window is missed, deals tend to be struck late, which requires correspondingly greater concessions to the proliferator.[2] But achieving these optimal early deals requires accurate forecasting and early detection of countries' nuclear intentions—a process we have little understanding of. Other prominent research argues that uncertainty about whether a country has or will soon invest in a nuclear weapons program can generate mistaken preventive wars, as in the ill-fated 2003 invasion of Iraq.[3] Yet we have little empirical evidence on how accurate such intelligence assessments are on average and therefore how uncertain policymakers should be.

While the ability of states to assess other countries' nuclear intentions may not tell us whether they can discern other sorts of intentions—for example, to go to war—it nonetheless is of clear policy relevance and establishes a baseline for assessing intentions to acquire powerful military technologies. Moreover, the nuclear domain presents many of the difficulties identified in the literature as making judgment of intentions especially

1. On the crucial role of intelligence assessments in nonproliferation, see Bas and Coe 2016; Long and Shifrinson 2019; Montgomery and Mount 2014; Nutt 2019; and Richelson 2007. On nonproliferation policies more broadly, see Braut-Hegghammer 2011; Coe and Vaynman 2015; Gavin 2015; Gerzhoy 2015; Gibbons 2020; Koch 2019; Kreps and Fuhrmann 2011; Mehta 2020; Miller 2014; Rabinowitz 2014; Solingen 2012; and Whitlark 2017.

2. Bas and Coe 2018.

3. Debs and Monteiro 2014. See Butt 2019 for a counterargument on the role of Iraq's perceived nuclear intentions in motivating the war.

International Organization 76, Spring 2022, pp. 487–507
© The IO Foundation 2021
doi:10.1017/S0020818321000345

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

challenging: there is a high degree of ambiguity, in part due to the dual-use nature of nuclear facilities; there are usually efforts at deception; and key decisions are often made by a small set of leaders, who can have widely divergent preferences. Finally, most of the assessments analyzed here are efforts to predict future intentions, which is generally considered harder than making judgments about the present.[4]

In the realm of nuclear proliferation, there have been four major works examining the accuracy of intelligence assessments. Yusuf surveyed proliferation predictions from US intelligence agencies and nongovernmental analysts since the onset of the Cold War, finding a significant bias toward alarmism and over-prediction of the extent of proliferation.[5] Looking at US estimates of countries with active nuclear weapons programs, Montgomery and Mount likewise find a poor track record, with a bias toward overestimation.[6] Bollfrass quantitatively examined the accuracy of US and European assessments during the Cold War. In contrast to the prior work, Bollfrass found a strikingly high accuracy rate for US and British assessments—over 90 percent—and a tendency toward underestimating rather than overestimating programs.[7] Most recently, Schneider qualitatively analyzed US assessments from the 1950s to 1990s and found a bias toward overestimating the number of countries that would acquire nuclear weapons, which he attributes to a focus on technological capacity rather than motives.[8]

The goal of this research note is to build on this literature by examining how assessment accuracy evolves over time. Specifically, I analyze close to 200 US assessments of countries' nuclear intentions, drawn from National Intelligence Estimates (NIEs) produced between 1957 and 1966. I find a high success rate overall, with close to 80 percent of testable assessments in the NIEs correctly differentiating between countries that would and would not proliferate. However, this number obscures wide variation in accuracy over time, as assessments improved from highly inaccurate in the late 1950s to highly accurate in the 1960s. Drawing on additional evidence and analysis, I conclude that learning from past failures played a major role in this improved performance.

This research note begins by describing the data used and the methodology. After analyzing the accuracy of US assessments overall, I explore why the assessments improved. I conclude with a discussion of the implications of the findings.

Data and Methods

Analysts in the US government have long sought to forecast the likelihood and consequences of the spread of nuclear weapons. In the late 1950s the intelligence

4. On the main barriers to accurately assessing intentions, see Jervis 2010; Mearsheimer 2001; Rosato 2014/2015; Rovner 2011; Yarhi-Milo 2013, 2014.
5. Yusuf 2009.
6. Montgomery and Mount 2014.
7. Bollfrass 2017.
8. Schneider 2020.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

community (IC) began producing NIEs that systematically tackled this problem, identifying which countries were currently seeking nuclear weapons and which were likely to do so over the next decade. The data for this research note are drawn from the NIEs that have been declassified, covering the period from 1957, when the first NIE was completed, to 1966, when basic agreement was reached between the United States and the Soviet Union on the Nonproliferation Treaty.[9]

This research builds on the prior work discussed but differs in a few key respects. Instead of examining estimates of the technical progress of a country's nuclear program and its likely date of acquisition—the subject of many of the estimates Montgomery and Mount and Schneider analyze—it zeroes in on estimates of a country's current and future *intentions* to develop nuclear weapons. This narrows the scope of inquiry but also broadens the aperture to include estimates for countries that never went on to seek nuclear weapons but were considered candidates for proliferation at the time. Thus the findings shed light on whether the IC was accurate in predicting who was likely to be *next* in seeking nuclear weapons, and the proportion of false positives—countries that were predicted to seek nuclear weapons but did not.

The analysis here also differs from Bollfrass's work in several ways. It restricts the analysis to the US IC, which has the most declassified estimates, enabling longitudinal analysis. Moreover, instead of treating the country-year as the unit of analysis and using a global sample, it uses the individual assessment of a given country, as explained later. This avoids an inflated sample size which might bias the analysis toward finding a high success rate and statistically significant results.[10] Finally, this research focuses on whether the quality of assessments changes over time, and also assesses more fine-grained process assessments—that is, not just whether a country would seek nuclear weapons but whether a given event would lead them to debate the nuclear option, or a particular arms control policy would affect their calculus.

I focus on NIEs, as opposed to the wide range of other estimates produced by the US government, for two main reasons. First, NIEs are widely regarded as the most important and influential. As one intelligence scholar puts it, "The NIE is arguably the highest form of intelligence art…The authors of these documents…are considered to be the best analysts in the US Intelligence Community."[11] According to Sherman Kent, who led the production of NIEs in the 1950s and 1960s, NIEs were intended "to command recognition and respect throughout the Government as the best available and presumably the most authoritative estimate."[12] NIEs were produced by the CIA, but only after extensive consultation with the government's other intelligence

9. Brands 2006, 108.

10. Bollfrass's data set includes countries that US NIEs do not explicitly mention and do not proliferate as "accurate" estimates. While this is true in a technical sense (since the estimates were supposed to identify all plausible proliferation risks), it gives credit to US analysts for accurately assessing countries with no realistic proliferation potential (Bollfrass 2017, 36–41).

11. Wheaton 2012, 330.

12. "The Law and Custom of the National Intelligence Estimate," in Kent 1994.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

agencies, whose leaders read and approved the final text.[13] Second, by examining a series of NIEs, rather than a wide range of estimates produced by different agencies and through different processes, we have a more homogeneous set of data, as NIEs followed mostly consistent procedures over time with respect to the drafting process.[14] This supports "apples to apples" comparisons across assessments and over time.

Coding the Accuracy of Assessments

The unit of observation is an individual assessment of a country's current or future intention to seek nuclear weapons.[15] I exclude assessments of countries that have already acquired nuclear weapons and leave out duplicate assessments from the same NIE. I omit vague assessments—words like "might," "could," and "is possible"—whose accuracy cannot be coded. I also exclude assessments that are conditional on an antecedent condition that did not occur—for example, "If Japan acquires nuclear weapons, South Korea will follow." Applying these coding rules to the seven declassified NIEs on proliferation under study—from 1957, 1958, 1960, 1961, 1963, 1964, and 1966—yields 199 distinct observations.[16]

One limitation of this study is that it relies on declassified assessments. Several of the NIEs have redacted portions, although these are relatively small compared to the length of the documents and may reflect information about sources and methods rather than the conclusions themselves. In any case, I draw on the assessments that are declassified as a first cut.

I assess the accuracy of the assessments in various ways, drawing on recent research on political forecasting.[17] First, I assess the *discrimination* of the assessments, or whether they accurately differentiate events that do occur from those that do not. Following Mandel and Barnes, I use a simple measure and code assessments as correctly classified if one of two conditions holds: greater than 50 percent odds are attached and the event does in fact occur in the specified time frame; or less than 50 percent odds are attached and the event does not occur in the specified time frame. Conversely, assessments are coded as incorrect if less then 50 percent odds are attached and the event does occur in the time frame, or greater than 50 percent odds are attached and the event does not occur in the time frame.[18]

13. "The Institutional Framework," in Kent 1994.
14. "The Making of an NIE," in Kent 1994.
15. In several cases, assessments were about groupings of countries, for example, whether European countries would jointly develop nuclear weapons. In a few other cases, assessments were global rather than country-specific—for example, "Other than X, no other countries are likely to seek nuclear weapons in the next ten years." I treat these cases as one observation.
16. NIEs were obtained from the National Security Archive, <http://nsarchive.gwu.edu/>.
17. Friedman 2019; Friedman, Lerner, and Zeckhauser 2017; Friedman and Zeckhauser 2016; Mandel and Barnes 2014; and Tetlock 2005.
18. It is possible, of course, that an analyst could be "correct" that an event has only a 30 percent chance of occurring, and it nonetheless occurs in a particular instance. Brier scores and the calibration index, discussed later, help account for this.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

Second, I code *false positives* and *false negatives.* A false negative is when an assessment predicted no proliferation but proliferation did in fact occur. A false positive is when an assessment predicted proliferation but it did not occur.

Third, I assess the overall accuracy of forecasts using Brier scores, which range between 0 and 1 and capture the difference between forecasted probabilities and actual outcomes.[19] This metric accounts not just for discrimination but also the *calibration* of the assessments: in other words, whether events forecasted as 75 percent likely, for instance, actually occur around 75 percent of the time. The further the forecasted probabilities are from the actual outcomes (meaning a less accurate forecast), the closer the Brier score is to 1. To calculate Brier scores, it is necessary to translate the qualitative probabilities in the NIEs into quantitative terms. I draw on research by Mandel, who conducted an experiment to gauge the numerical probabilities people associate with a range of linguistic probability terms.[20]

Finally, I assess calibration separately by calculating a calibration index, which ranges between 0 and 1, where 0 is perfect calibration. This statistic captures the mean square differences between the forecasted probabilities and the actual frequency of events, averaged over each of the probability categories and weighted by the number of observations in each category.[21]

Coding whether an NIE accurately assessed that a country would seek nuclear weapons poses a few challenges. It can be ambiguous when the weapons program started or stopped. Moreover, the dividing line between exploration and a determined pursuit of nuclear weapons can be blurry.[22] When the language in NIEs refers to the outcome of interest for this study, deciding to seek nuclear weapons, I rely on Singh and Way's coding, which draws on historical sources to cover all countries from 1945 to 2009 and capture instances where there is a "political decision by cabinet-level officials, movement toward weaponization, or development of single-use, dedicated technology."[23] As a robustness check, I use Bleek's coding of nuclear pursuit, which uses the same definition as Singh and Way but draws on different historical sources.[24] When it comes to more fine-grained *process* assessments—for example, whether a country would seek to keep its nuclear options open, debate developing nuclear weapons, seek nuclear weapons from a third country, or whether a particular US policy would sway their decisions—I rely on secondary sources for countries' nuclear histories. About 45 percent of the assessments in the data set are about process.

Another potential concern is that the process and outcome assessments are not always independent of each other and that this could artificially inflate the accuracy rate. For example, one assessment may predict that India will seek nuclear weapons,

19. Specifically, the Brier score is the mean squared error of the forecast in the sample.

20. Mandel 2015; also see Friedman and Zeckhauser 2016. There were six assessments where I was unable to translate a qualitative assessment into a quantitative one.

21. See Mandel and Barnes 2014, 10986; and Tetlock 2005, 275.

22. For example, see Montgomery and Sagan 2009.

23. Singh and Way 2004, 866. For updated codings, see Way and Weeks 2014, appendix.

24. See Bleek 2017.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

and another in the same year may predict a specific event that will trigger this decision and how India is likely to respond to a US nonproliferation initiative. Logically, if the former assessment is correct, the latter are more likely to be as well. To address this concern, I divide the sample into process and outcome observations and assess whether including process assessments drives up the accuracy rate.[25]

A final challenge in determining accuracy is the possibility of self-defeating prophecies, whereby the assessors are correct that a country *would* seek nuclear weapons but it was dissuaded by successful policy interventions. To some extent, this problem is insurmountable, since we cannot test the alternate history where a given policy intervention did not take place. On the other hand, the assessors were sensitive to this issue and explicitly addressed the possibility that US or other nonproliferation efforts would affect a state's calculus. Thus, if an NIE predicted that US policies would affect a state's choices, I code this as accurate if the policy occurred and historical evidence suggests it swayed the country's nuclear decisions. In other words, the assessments under study include those that predict the impact of US policies.

Analysis and Results

Pooled Sample

So, how accurate were US assessments of other countries' nuclear intentions? Starting with discrimination, 78 percent were correctly classified, which seems decent if one's prior is that it is highly challenging to judge current or future intentions.[26] This rate is lower, however, than the over 90 percent recorded in Bollfrass's research, or the 94 percent found in Mandel and Barnes's recent analysis of Canadian intelligence estimates for the Middle East and Africa.[27] If we exclude observations about process and look only at outcomes, accuracy increases to 83.3 percent. This suggests that including process assessments is actually driving *down* the accuracy rate rather than leading to an artificially high rate.

But is 78 percent actually impressive, given the type of events being forecast? After all, if seeking nuclear weapons is a rare event, and assessors always (or almost always) predicted no proliferation, they would have a high accuracy rate but also be unhelpful in predicting the rare cases where proliferation does occur. However, proliferation is not a rare outcome in this data set, since the assessments focus on countries deemed to be relatively high proliferation risks. And predictions of proliferation and nonproliferation were both reasonably accurate (Table 1). While predictions that countries *would not* seek (or move toward) nuclear weapons were better (85.3 percent), predictions that they *would* were still mostly on target (68.7 percent). Indeed, if an assessor simply

25. Process assessments are those whose accuracy required looking at sources other than Singh and Way's coding of nuclear weapons programs.

26. For the success rate by country, see the online supplement.

27. Bollfrass 2017; and Mandel and Barnes 2014.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

predicted no proliferation in every instance—the modal outcome—they would only have been correct 63 percent of the time, as opposed to the 78 percent we actually observe using the NIE data, a statistically significant difference.

TABLE 1. *Discrimination of intelligence assessments*

	Observations	*Correctly classified*
Proliferation predictions	83	68.7%
Nonproliferation predictions	116	85.3%
Total	199	78.4%

A stronger counterargument to this optimistic take is that the analysis excludes vague assessments, whose accuracy could not be judged. This probably biases in favor of finding a high success rate, since vague estimates could indicate a lack of confidence. Indeed, about 20 percent of assessments were too vague to be included in the data set, which highlights an important limitation to the utility of the assessments.

With respect to false positives versus false negatives, of the forty-three incorrect assessments, 60.4 percent were false positives and 39.6 percent were false negatives. This ratio, about 1.5 to 1, is what one might expect in a situation where the costs of inaction (failing to prevent proliferation) are perceived to be high, thus incentivizing false positives.[28]

Turning to the Brier score, which helps account for both discrimination and calibration, the result is 0.176. This is far from a perfect score of 0 (no error in forecasts) but better than the 0.23 score of predicting the base rate probability of proliferation in each instance, a difference that is statistically significant at the 95 percent level.

The calibration index for the full sample is 0.026, which is not as impressive as the results of previously published analyses of expert forecasts in a variety of different issue areas.[29] In the online supplement (Table A7), I provide raw data on calibration, that is, the observed frequency of events for each forecasted probability. These data show that the forecasts tended to be overconfident. For example, events that were deemed to have no chance of occurring actually occurred about 8 percent of the time, and events forecast as certain only occurred about 78 percent of the time.

Change over Time

These results suggest that the NIEs' assessments of intentions were reasonably accurate as a whole, though by no means perfect. But how did this change over time? Using the basic measure of discrimination, Figure 1 breaks up the sample by year and

28. Gartner 2013.
29. Stasny and Lehner 2018, 209–10.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

reveals a striking improvement between the initial assessments in 1957–58 and the later assessments of 1960–66, a statistically significant difference. Figure 2 demonstrates that this change is driven by a downward trend in the rate of false positives, while there is no clear trend in the rate of false negatives. The difference between the 1957–58 and 1960–66 periods in the rate of false positives is statistically significant; the difference in the rate of false negatives is not.

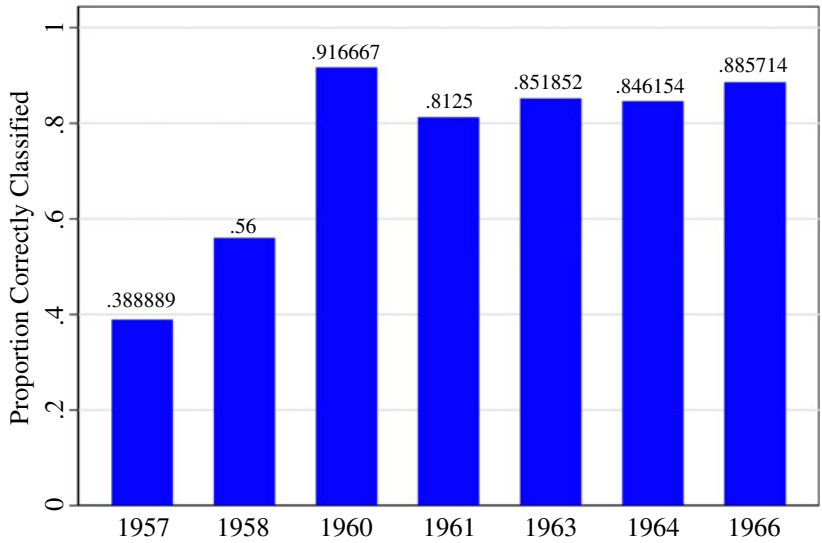


FIGURE 1. *Performance by year*

Table 2 shows Brier scores calculated separately for the 1957–58 and 1960–66 periods, again demonstrating a dramatic improvement between the 1950s and 1960s. The score for 1957–58 is worse than what would be achieved by predicting the base rate probability of proliferation for this period. Starting in the 1960s, this trend reverses, with Brier scores showing that the NIEs outperform the predictions using base rates. The difference in scores between the 1957–58 and 1960–66 periods is statistically significant at the 95 percent level.

Calibration likewise dramatically improved across the two periods, from 0.189 in the 1950s to 0.012 in the 1960s. The 0.012 figure for the later period is on par with the performance of expert analysts in other published studies of forecasting.[30] The raw data on calibration (Tables A8 and A9) tell a similar story. In the 1950s, observed frequencies were way off the mark of forecast probabilities, and the analysts had a striking bias toward incorrectly estimating events as highly likely. For instance, of

30. Stasny and Lehner 2018, 209–10.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

events forecast with 75 percent probability, only 35 percent occurred; of events forecast to be certain, only 43 percent occurred. In the 1960s, calibration dramatically improved. For example, the events characterized as 75 percent likely occurred 76 percent of the time in actuality, and "certain" predictions came to pass 94 percent of the time.

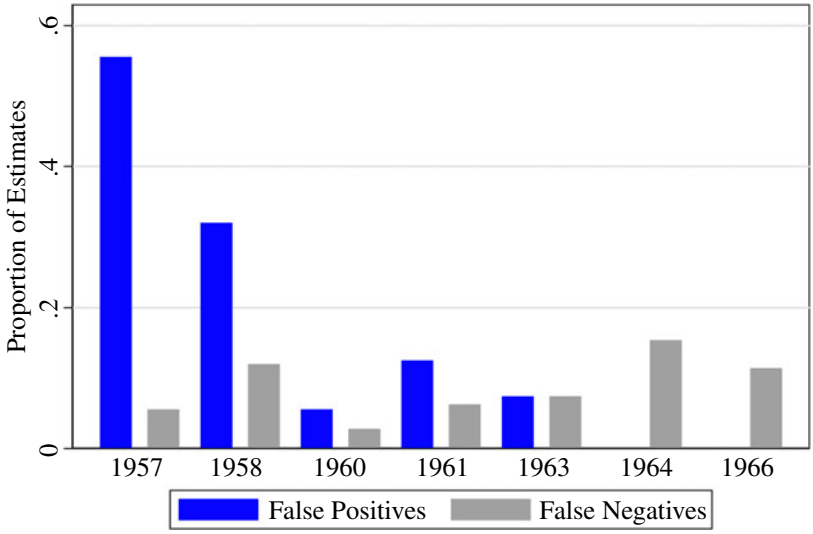


FIGURE 2. *Error type, by year*

TABLE 2. *Brier scores over time*

Period	*n*	*Brier score*	*Score with base rate*
1957–58	43	0.381	0.243
1960–66	156	0.119	0.230

In the online supplement, I show that the improvement over time is robust to accounting for the composition of countries and measurable characteristics like regime type and relation to the United States, as well as substituting Bleek's coding of nuclear programs. To test whether the results still hold after the 1960s, I coded a similar 1974 NIE and find the improvement was sustained. The results also hold if we restrict the data set to outcome assessments.

Accounting for the Improved Accuracy

The evidence demonstrates that estimates of proliferation intentions significantly improved over time—in particular by reducing the rate of false positives—but it

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

does not demonstrate *how* or *why* this improvement occurred. One possibility is that the improvement was the result of learning from the mistakes of early assessments. Jack Levy defines learning as "a change of beliefs (or the degree of confidence in one's beliefs) or the development of new beliefs, skills, or procedures as a result of the observation and interpretation of experience." Learning "is not a passive activity" and involves effort to gather the information needed to evaluate ideas, as well as trial-and-error policy changes to assess whether outcomes improve. Learning can be either causal, that is, changing beliefs about cause and effect, or diagnostic, that is, changing beliefs "about the definition of the situation or the preferences, intentions, or relative capabilities of others." As a general rule, learning occurs at the individual level; organizations only "learn" when individuals take "action to change organizational procedures, which leads to a change in organizational behavior."[31]

Using this definition, if learning is responsible for the IC's improved performance, we would expect to find evidence that (1) the IC gathered information to assess its performance; (2) this information led to a change in skills, procedures, or beliefs that was institutionalized at the organizational level; and (3) these changes led to the improved performance we observe.

On the first score, declassified documents make clear that the IC was gathering information on the accuracy of its NIEs. In September 1958, two months after that year's NIE was completed, the US embassy in Sweden cabled Washington to register its dissent from how the estimate had characterized Sweden's nuclear intentions, arguing that it was "not supported by present facts" and recommending that the estimate lower the likelihood of Swedish proliferation.[32] The following month, the State Department cabled its embassies in Western Europe to ask for more information on French and possible cooperative European nuclear weapons programs, which it noted had been a weak point in the 1958 NIE. Rather than approaching French or European officials directly on this question, it suggested indirect approaches, as well as seeking to glean insight from third parties.[33]

A few weeks later, a postmortem on the 1958 NIE was produced, expressly to assess its quality. The report identified improvements over the 1957 installment, but nonetheless judged that "in many respects the available information did not permit confident judgments of the timing and extent of future programs, to the degree that we believe information could be obtained with a higher priority effort."[34] As targets for greater collection going forward, the report singled out France, Italy, West Germany, and other European nations, noting that "it should be possible to increase the scale of appropriate operation to obtain fuller and more accurate information." In a separate recommendation, the postmortem called for more

31. Levy 1994, 283–88.
32. National Security Archive (NSA) 2005, doc. 4a.
33. Ibid., doc. 4b.
34. A September 1960 validity study on the 1958 NIE noted that the "estimate probably over-estimated the likelihood of certain European countries initiating and carrying out independent nuclear weapons programs" (Central Intelligence Agency 1960).

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

reporting on the "the political factors in the possible decisions of fourth countries to initiate nuclear weapons programs," which "could take greater account of the underlying forces, often not openly expressed, that may influence the making and shape of such decision."[35]

The postmortem thus suggests that at least two types of learning were taking place: development of new procedures, such as gathering more information and collating information already available from different sources for a specific set of Western European countries; and possibly a change in causal beliefs, highlighting the importance of political factors in affecting states' nuclear decisions.

There is also evidence of an effort to institutionalize this learning. In addition to the memo sent to Western European embassies in September 1958, the entire postmortem was transmitted in December 1958 to the US embassies in Switzerland, West Germany, Belgium, the Netherlands, the United Kingdom, Luxembourg, France, Italy, Sweden, and Japan. In delivering the report, the State Department again emphasized the importance of political factors, both domestic and international: "Analysis of political factors, such as, for example, the interplay of party orientation and rivalries, the effect of scheduled elections, the implications of national aims and roles, and a country's relations with and attitudes toward other powers is of material importance to an assessment of 'fourth power' intentions in the nuclear field."[36] As later noted in a 1961 postmortem on Israel's nuclear program, these recommendations were *not* communicated to *all* US embassies abroad, possibly contributing to US failure to detect Israel's nuclear program in a timely fashion.[37] But it wasn't just US embassies that were receiving new instructions. The postmortem also called on the Atomic Energy Commission and the Joint Atomic Energy Intelligence Committee to work on gathering more information, and directed the US Intelligence Board, which helped oversee the production of NIEs, to ensure greater emphasis on political factors.[38]

But did this actually lead to the improvement we observe? If so, we would expect to find that assessments improved *especially* for countries that were identified as high-priority targets, where more information and effort were requested, and less so for other countries. And if learning is a result not just of gathering more information but also of the IC changing its type of analysis, we should also expect to find that *initial* assessments of countries were more accurate after 1958. If this latter implication holds, it helps rule out the possibility that the improved performance was entirely due to having observed a given country's behavior for longer, as opposed to new information-gathering procedures or a focus on different causal factors.

The data confirm both observable implications. In countries identified as targets in 1958, there is a massive improvement (Figure 3). Brier scores show the same pattern,

35. NSA 2005, doc. 4c.
36. Ibid.
37. US Joint Atomic Energy Intelligence Committee 1961.
38. NSA 2005, doc. 4c.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

with assessments of countries designated as targets changing from highly uninformative in the 1950s to highly informative in the 1960s (Table A5). For countries not identified as high-priority targets, there is only very modest, nonsignificant improvement after 1958 (Figure 4). Regression models that account for country fixed effects and potential confounders and use the basic measure of discrimination as the dependent variable similarly show significantly better performance in the 1960s for countries identified as targets and nonsignificant improvement for other countries (Table A6).

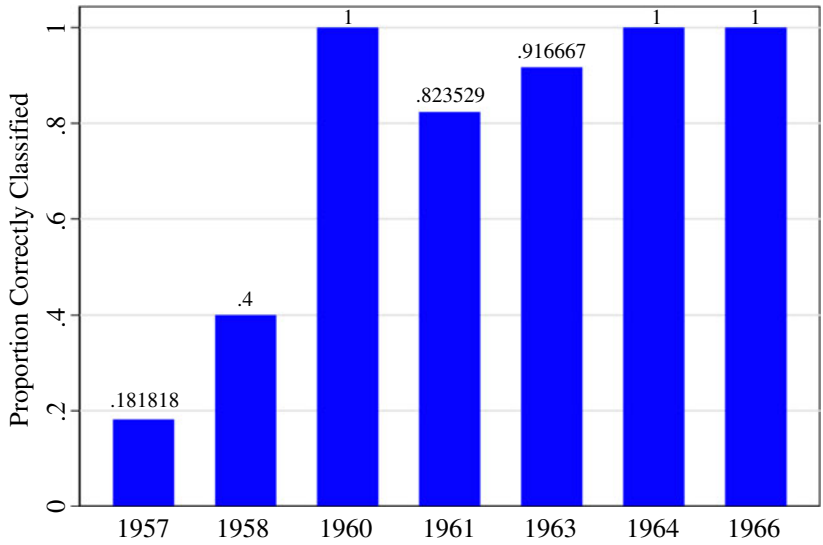


FIGURE 3. *Performance by year, intel targets*

Finally, if we look only at countries that are being assessed for the first time, accuracy also improved dramatically—a statistically significant difference (Figure 5). This suggests that the IC was successful not only at improving accuracy with specific target countries but also at improving its performance with new countries going forward.

Taken together, the qualitative and quantitative evidence strongly suggests that the IC learned from its early failures and took effective steps to improve assessments. To increase confidence in this finding, I turn to a short case study of Sweden to tease out how exactly assessments improved after 1958.

The Swedish Case

One case that illustrates the improvement of US assessments is Sweden, one of the countries where US officials were directed to step up their collection and analysis efforts after 1958. In both the 1957 and 1958 NIEs, the IC was overly alarmist in

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

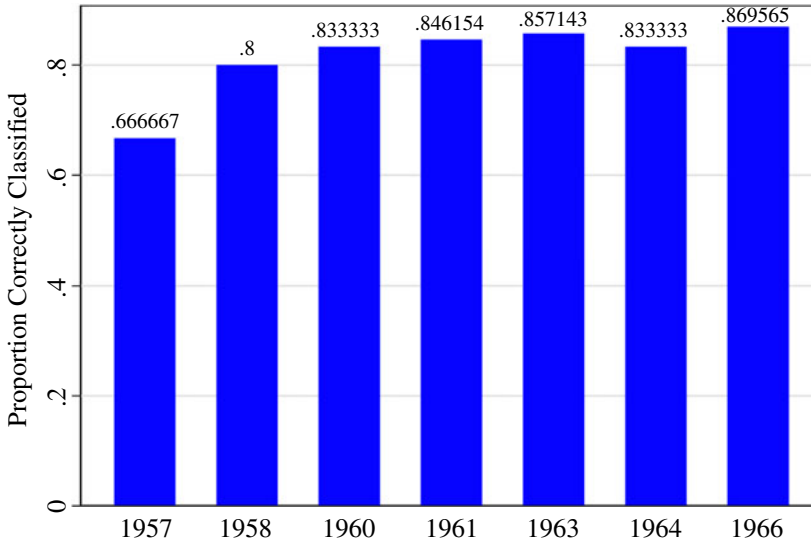


FIGURE 4. *Performance by year, non-targets*

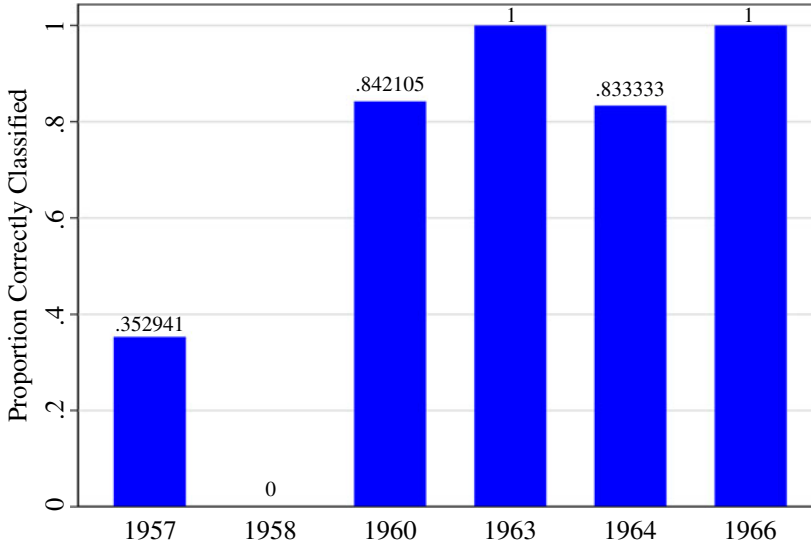


FIGURE 5. *Accuracy of first assessments*

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

predicting that Stockholm would develop nuclear weapons. According to the 1957 NIE, "Although a vocal minority opposes equipping Sweden's armed forces with nuclear weapons, public opinion in general appears to support the government view that nuclear weapons are essential to Sweden's defense." The NIE concluded that "in the absence of substantial progress in disarmament, Sweden will initiate production of nuclear weapons as the necessary materials become available."[39] The following year's NIE came to an identical conclusion based on slightly different logic and evidence. The report acknowledged that in Sweden there was "strong opposition to nuclear weapons on moral grounds" as well as "uncertainty as to whether Sweden's traditional neutrality is better served by foregoing or by producing atomic arms," given that developing nuclear weapons may in fact trigger an attack. Nevertheless, the NIE judged, "the major political parties and most leaders are inclined to compromise [on] this issue by supporting the manufacture of tactical and air defense atomic weapons, while avoiding—at least initially—the more 'provocative' longer range missiles."[40]

Based on what we know now, there are several things the 1957 and 1958 NIEs got right, even though they erred in predicting that Sweden would ultimately go nuclear. They were correct that a plurality of the public supported acquiring nuclear weapons and that there was also strongly felt opposition, in segments of both the populace and the elite. They were also right that these divisions would incline the Swedish government to seek a compromise.[41] However, the estimates exaggerated the support for nuclear weapons in the political leadership and got wrong the form that this compromise would take.

In July 1958, the Swedish parliament declined to fund a program to develop nuclear weapons, instead authorizing "protection research," which allowed research into how to defend against nuclear weapons.[42] A few months later, prime minister Tage Erlander of the Social Democratic Party (SDP) established a committee to study the issue and help resolve the debate. The committee, which included strong opponents of nuclear weapons, recommended that the government continue to follow the line set out by parliament: that is, no development of nuclear weapons but continued research into defense against them, which would allow Sweden to advance its nuclear expertise without taking the provocative step of building the bomb.[43] As Arnett explains, "Erlander claimed to be sympathetic with the pronuclear side, but did not want to risk dividing the party over the issue." The antinuclear position was strengthened by the outcome of the 1958 election, which compelled the SDP into an alliance with the strongly antinuclear Communist Party.[44] Cole

39. NSA 2005, doc. 2.
40. Ibid., doc. 3a.
41. Arnett 1998, 35–38; Cole 1994, 94–99; and Jonter 2001, 42–49.
42. Jonter 2001, 43–44.
43. Ibid.; and Arnett 1998, 36.
44. Arnett 1998, 36.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

likewise concurs that the SDP "sought to avoid a split that would seriously weaken the party," leading to a decision to defer producing nuclear weapons.[45]

As noted before, the US embassy in Sweden had criticized the 1958 NIE as alarmist. It specifically took issue with the idea that Swedish leaders would "compromise" by producing short-range nuclear weapons, arguing that this was based on a misreading of Swedish politics. According to the memo:

> Of the four non-Communist parties, only the leadership of the opposition Conservative Party has pleaded the case for atomic weapons as a modern defensive necessity while the other three parties, including the governing Social Democrats, have not yet taken definitive positions and are not likely to do so in the near future. In these circumstances the distinction between tactical and strategic atomic weapons has not been of much significance as of yet. The Embassy therefore recommends that the sentence be revised to read as follows: "The question of domestic production of these weapons is still a matter of party and public debate, with party opinion apparently lagging somewhat behind some segments of public opinion in favor of Sweden's acquiring them."

The memo recommended that the NIE's prediction for Sweden be revised to include more uncertainty: Sweden would "probably" produce nuclear weapons in the absence of disarmament measures, as opposed to Sweden "will" do so.[46]

By 1960, the US assessment of Sweden's intentions had changed, in line with the feedback from the embassy and the 1958 postmortem's recommendation to focus more on political factors. The 1960 NIE especially emphasized domestic politics as a major inhibitor of any decision to develop nuclear weapons, describing "strong opposition within the SDP," and noting that "many Swedes, and particularly those in the Social-Democratic leadership, continue to hope that the international climate will develop so as to obviate the need to make a decision." Whereas the previous NIEs had estimated that "substantial progress in disarmament" would be needed to deter Swedish proliferation, by 1960 the IC had determined that "a nuclear test ban or positive steps toward disarmament—*or even a reasonable hope that one or the other would soon come about*—would probably be sufficient."[47] This assessment was right on the mark. Erlander wanted to avoid tearing his party apart over the nuclear issue. Moreover, beyond domestic politics, the Social Democrats felt the international situation gave them the latitude to wait and see. As Jonter describes, "The committee of the Social Democratic Party Council took the view that the future security policy situation should determine whether or not Sweden should acquire nuclear weapons," and that "Sweden had a breathing space

45. Cole 1994, 97.
46. NSA 2005, doc. 4a.
47. NSA 2005, doc. 5. Emphasis added.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

at least until the mid-1960s, when international developments should provide an answer."[48]

A 1961 NIE came to similar conclusions; the 1963 NIE was vaguer in its predictions—"whether the Swedes will actually go on to produce nuclear weapons is uncertain;"[49] and the 1964 NIE returned to domestic politics as a key inhibiting factor:

> Swedish capabilities for nuclear weapons production have increased in the past year, but public opposition to a weapons program probably has also increased. The view is widely held in Sweden, and has increased in strength since the signing of the partial test ban treaty, that nothing should be done to encourage proliferation of nuclear weapons, but that Sweden instead should use its influence to get the present nuclear powers to agree on further controls over these weapons.

Moreover, the ruling SDP was "on record as wishing to postpone a decision on nuclear weapons as long as possible." For all these reasons, the NIE concluded, "the chances that the Swedish Government will decide upon a nuclear weapons program are less than even during the next decade."[50] Two years later, the NIE highlighted similar factors—public opinion and the government's desire for international arms control—as making it likely that Sweden would "probably continue to postpone a decision for the next several years."[51]

These assessments from the mid-1960s were on point, accurately focusing on domestic opposition and the status of international arms control. A 1962 government study, for example, concluded that "international developments must determine a decision for or against in Sweden," namely the success of arms control efforts.[52] Sweden welcomed and signed the 1963 Partial Test Ban Treaty, and a faction of Social Democrats hoped that this "would be the first of several arms control measures that would permit Sweden to reject nuclear weapons" once and for all. Nonetheless, as accurately judged by the NIEs, Swedish leaders continued to defer a final decision.[53] With respect to domestic politics, by the early 1960s public opinion had decisively turned against Swedish acquisition of nuclear weapons.[54] As Erlander later explained in his memoirs, his government ultimately decided against nuclear weapons for both international and domestic reasons. The prime minister "did not want to divide the party or destabilize the policy of non-alignment, issues of domestic politics directly related to Swedish regulatory and constitutive norms." Moreover, Erlander wanted to arrest nuclear proliferation internationally and believed Sweden would not use nuclear weapons; therefore, it would be better off if it would

48. Jonter 2001, 44.
49. NSA 2005, doc. 6b, doc. 8.
50. NSA 2012, doc. 3.
51. NSA 2005, doc. 12.
52. Jonter 2001, 55.
53. Cole 1994, 99.
54. Ibid., 94.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

"campaign instead for complete nuclear disarmament."[55] These factors ultimately combined to lead Sweden to renounce nuclear weapons and sign the Nonproliferation Treaty in 1968. In line with the quantitative findings earlier, US assessments of Sweden significantly improved after 1958. And they improved in the way the 1958 postmortem called for: by providing more detailed information and more focus on political rather than technical dynamics.[56]

Alternative Explanations

In the online supplement, I assess a variety of alternative explanations for improvement over time—including changes in technology, changes in the nonproliferation regime, and turnover in who is doing the assessments—and find that none are as persuasive as learning. I also assess the quality of reasoning in the assessments in two additional cases and find that the analysis was mostly right for the right reasons.

Conclusion

This research note has offered new evidence on US assessments of foreign nuclear intentions. The results suggest that testable US assessments were quite accurate overall and dramatically improved over the period studied. A combination of qualitative and quantitative evidence suggests that the improvement in performance was largely driven by a process of learning, whereby the IC recognized past failures in assessments, took steps to address them, and subsequently performed better.

These findings have important implications for the study of intelligence and nuclear proliferation. With respect to intelligence, the results cast doubt on the notion that assessing the intentions of foreign countries is a fool's errand and suggest that at least in this issue area, intelligence agencies are capable of learning the "right" lessons rather than reflexively maladapting and flipping to the opposite errors from before—a pathology often discussed in the intelligence literature.[57] They also should help temper the pessimism expressed by many practitioners and scholars of intelligence analysis, as in Lowenthal's claim that "No one has yet come up with any methodologies, machines, or thought processes that will appreciably raise the Intelligence Community's" accuracy.[58] Indeed, the findings

55. Arnett 1998, 37.

56. Ideally, we could also gather data on the source material going into NIEs to assess whether the improved accuracy was related to focusing on particular types of sources or targets—that is, human intelligence related to politicians' views, or those of scientists and military officials, and/or intercepted signals. Unfortunately, information on sources and methods is absent or redacted from the NIEs.

57. Jervis 2006, 28–29; and Tetlock and Mellers 2011, 543.

58. Lowenthal 2008, 314.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

support the idea that systematic reviews of intelligence forecasts may improve rather than erode the credibility of intelligence agencies.[59] They also show that learning takes place in two ways: changing organizational procedures to gather more information in key countries, and changing the analysis to give greater weight to political rather than technical factors.

The findings also deepen our understanding of the nonproliferation process. The prevalence of nonproliferation bargains early in the proliferation process has been established.[60] The findings in this paper suggest that accurate US intelligence assessments of nuclear intentions may be partly responsible for success in containing nuclear programs early, and also that this accuracy was not automatic but evolved over time. The findings also imply that US decision makers perhaps should not be so uncertain about intelligence assessments of countries' nuclear intentions when deciding whether to go to war or implement other nonproliferation tools. While it is not perfect, the overall track record is quite impressive in the assessments under study here. The failure of the IC in the case of Iraq should not blind us to its broader track record of success.

Data Availability Statement

Replication files for this research note may be found at <https://doi.org/10.7910/DVN/BI4FJW>.

Supplementary Material

Supplementary material for this research note is available at <https://doi.org/10.1017/S0020818321000345>.

References

Arnett, Eric. 1998. Norms and Nuclear Proliferation: Sweden's Lessons for Assessing Iran. *Nonproliferation Review* 5 (2):32–43.

Bas, Muhammet, and Andrew Coe. 2016. A Dynamic Theory of Nuclear Proliferation and Preventive War. *International Organization* 70 (4):655–85.

Bas, Muhammet, and Andrew Coe. 2018. Give Peace a (Second) Chance: A Theory of Nonproliferation Deals. *International Studies Quarterly* 62 (3):606–17.

Bleek, Philipp. 2017. When Did (and Didn't) States Proliferate? Chronicling the Spread of Nuclear Weapons. Discussion Paper, Project on Managing the Atom, Belfer Center for Science and International Affairs, Harvard Kennedy School.

Bollfrass, Alexander. 2017. The Half-Lives of Others: The Democratic Advantage in Nuclear Intelligence Assessment. PhD diss., Princeton University.

59. Friedman 2019, chap. 5; and Friedman and Zeckhauser 2016, 195.
60. Bas and Coe 2018.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

Brands, Hal. 2006. Rethinking Nonproliferation: LBJ, the Gilpatric Committee, and US National Security Policy. *Journal of Cold War Studies* 8 (2):83–113.
Braut-Hegghammer, Malfrid. 2011. Revisiting Osirak: Preventive Attacks and Nuclear Proliferation Risks. *International Security* 36 (1):101–32.
Butt, Ahsan. 2019. Why Did the United States Invade Iraq in 2003? *Security Studies* 28 (2):250–85.
Central Intelligence Agency. 1960. Validity Study on NIE 100-2-58: Development of Nuclear Capabilities by Fourth Countries: Likelihood and Consequences, Dated 1 July 1958. Available at <https://www.cia.gov/readingroom/document/cia-rdp82m00097r000600020014-9>.
Coe, Andrew, and Jane Vaynman. 2015. Collusion and the Nuclear Nonproliferation Regime. *Journal of Politics* 77 (4):983–97.
Cole, Paul. 1994. *Sweden Without the Bomb: The Conduct of a Nuclear-Capable Nation Without Nuclear Weapons*. RAND.
Debs, Alexandre, and Nuno Monteiro. 2014. Known Unknowns: Power Shifts, Uncertainty, and War. *International Organization* 68 (1):1–31.
Friedman, Jeffrey. 2019. *War and Chance: Assessing Uncertainty in International Politics*. Oxford University Press.
Friedman, Jeffrey, Jennifer Lerner, and Richard Zeckhauser. 2017. Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Professionals. *International Organization* 71 (4):803–26.
Friedman, Jeffrey, and Richard Zeckhauser. 2016. Why Assessing Estimative Accuracy Is Feasible and Desirable. *Intelligence and National Security* 31 (2):178–200.
Gartner, Scott Sigmund. 2013. All Mistakes Are Not Equal: Intelligence Errors and National Security. *Intelligence and National Security* 28 (5):634–54.
Gavin, Francis. 2015. Strategies of Inhibition: US Grand Strategy, the Nuclear Revolution, and Nonproliferation. *International Security* 40 (1):9–46.
Gerzhoy, Gene. 2015. Alliance Coercion and Nuclear Restraint: How the United States Thwarted West Germany's Nuclear Ambitions. *International Security* 39 (4):91–129.
Gibbons, Rebecca Davis. 2020. Supply to Deny: The Benefits of Nuclear Assistance for Nuclear Nonproliferation. *Journal of Global Security Studies* 5 (2):282–98.
Jervis, Robert. 2006. Reports, Politics, and Intelligence Failures: The Case of Iraq. *Journal of Strategic Studies* 29 (1):3–52.
Jervis, Robert. 2010. *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War*. Cornell University Press.
Jonter, Thomas. 2001. Sweden and the Bomb: The Swedish Plans to Acquire Nuclear Weapons, 1945–1972. *SKI Report* 01:33.
Kent, Sherman. 1994. *Sherman Kent and the Board of National Estimates: Collected Essays*. CIA Center for the Study of Intelligence, Washington, DC. Available at <https://www.cia.gov/static/c04dcfa20465147ef-fabe492a1df4612/sherman-kent-and-the-board-of-national-estimates-collected-essays.pdf>.
Koch, Lisa. 2019. Frustration and Delay: The Secondary Effects of Supply-Side Proliferation Controls. *Security Studies* 28 (4):773–806.
Kreps, Sarah, and Matthew Fuhrmann. 2011. Attacking the Atom: Does Bombing Nuclear Facilities Affect Proliferation? *Journal of Strategic Studies* 34 (2):161–87.
Levy, Jack. 1994. Learning and Foreign Policy: Sweeping a Conceptual Minefield. *International Organization* 48 (2):279–312.
Long, Austin, and Joshua Shifrinson. 2019. How Long Until Midnight? Intelligence-Policy Relations and the United States Response to the Israeli Nuclear Program, 1959–1985. *Journal of Strategic Studies* 42 (1):55–90.
Lowenthal, Mark. 2008. Towards a Reasonable Standard for Analysis: How Right, How Often on Which Issues? *Intelligence and National Security* 23 (3):303–15.
Mandel, David. 2015. Accuracy of Intelligence Forecasts from the Intelligence Consumer's Perspective. *Policy Insights from the Behavioral and Brain Sciences* 2 (1):111–20.
Mandel, David, and Alan Barnes. 2014. Accuracy of Forecasts in Strategic Intelligence. *Proceedings of the National Academy of Sciences* 111 (30):10984–89.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

Mearsheimer, John. 2001. *The Tragedy of Great Power Politics*. Norton.
Mehta, Rupal. 2020. *Delaying Doomsday: The Politics of Nuclear Reversal*. Oxford University Press.
Miller, Nicholas L. 2014. The Secret Success of Nonproliferation Sanctions. *International Organization* 68 (4):913–44.
Montgomery, Alexander, and Adam Mount. 2014. Misestimation: Explaining US Failures to Predict Nuclear Weapons Programs. *Intelligence and National Security* 29 (3):357–86.
Montgomery, Alexander, and Scott Sagan. 2009. The Perils of Predicting Proliferation. *Journal of Conflict Resolution* 53 (2):302–28.
National Security Archive. 2005. National Intelligence Estimates of the Nuclear Proliferation Problem. Electronic Briefing Book 155, edited by William Burr. Available at <https://nsarchive2.gwu.edu//NSAEBB/NSAEBB155/>.
National Security Archive. 2012. Declassified 1964 National Intelligence Estimate Predicts India's Bomb but Not Israel's. Electronic Briefing Book 401, edited by William Burr. Available at <https://nsarchive2.gwu.edu/nukevault/ebb401/>.
Nutt, Cullen. 2019. Proof of the Bomb: The Influence of Previous Failure on Intelligence Judgments of Nuclear Programs. *Security Studies* 28 (2):321–59.
Rabinowitz, Or. 2014. *Bargaining on Nuclear Tests: Washington and Its Cold War Deals*. Oxford University Press.
Richelson, Jeffrey. 2007. *Spying on the Bomb: American Nuclear Intelligence from Nazi Germany to Iran and North Korea*. Norton.
Rosato, Sebastian. 2014/2015. The Inscrutable Intentions of Great Powers. *International Security* 39 (3): 48–88.
Rovner, Joshua. 2011. *Fixing the Facts: National Security and the Politics of Intelligence*. Cornell University Press.
Schneider, Jonas. 2020. Predicting Nuclear Weapons Proliferation. In *The Politics and Science of Prevision*, edited by Andreas Wenger, Ursula Jasper, and Myriam Dunn Cavelty, 194–212. Routledge.
Singh, Sonali and Christopher Way. 2004. The Correlates of Nuclear Proliferation: A Quantitative Test. *Journal of Conflict Resolution* 48 (6):859–85.
Solingen, Etel, ed. 2012. *Sanctions, Statecraft, and Nuclear Proliferation*. Cambridge University Press.
Stasny, Bradley, and Paul Lehner. 2018. Comparative Evaluation of the Forecast Accuracy of Analysis Reports and a Prediction Market. *Judgment and Decision Making* 13 (2):202–11.
Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.
Tetlock, Philip, and Barbara Mellers. 2011. Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong. *American Psychologist* 66 (6):542–54.
US Joint Atomic Energy Intelligence Committee. 1961. Post-Mortem on SNIE 100-8-60: Implications of the Acquisition by Israel of a Nuclear Weapons Capability, 31 January. Available at <http://digitalarchive.wilsoncenter.org/document/121959>.
Way, Christopher, and Jessica Weeks. 2014. Making It Personal: Regime Type and Nuclear Proliferation. *American Journal of Political Science* 58 (3):705–19.
Wheaton, Kristan. 2012. The Revolution Begins on Page Five: The Changing Nature of NIEs. *International Journal of Intelligence and Counterintelligence* 25 (2):330–49.
Whitlark, Rachel. 2017. Nuclear Beliefs: A Leader-Focused Theory of Counter-Proliferation. *Security Studies* 26 (4):545–74.
Yarhi-Milo, Keren. 2013. In the Eye of the Beholder: How Leaders and Intelligence Communities Assess the Intentions of Adversaries. *International Security* 38 (1):7–51.
Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton University Press.
Yusuf, Mooed. 2009. Predicting Proliferation: The History of the Future of Nuclear Weapons. Policy Paper 11, Brookings Institution.

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press

Author

Nicholas L. Miller is Associate Professor in the Department of Government, Dartmouth College. He can be reached at nicholas.l.miller@dartmouth.edu.

Acknowledgments

I thank Mark Bell, Alex Bollfrass, Jeffrey Friedman, Guy Laron, Julia Macdonald, Rolff Mowatt-Larsen, Jeremi Suri, Ben Valentino, Arne Westad, participants in Tobin Project workshops, and the editors and reviewers for their helpful comments and feedback.

Key Words

Nuclear weapons; nuclear proliferation; intelligence; intentions; learning

Date received: December 6, 2018; Date accepted: February 19, 2021

https://doi.org/10.1017/S0020818321000345 Published online by Cambridge University Press