Online Appendix for "Learning to Predict Proliferation."

Table of Contents | Page

Performance by Country Assessed

Table A1

Country/Grouping	# of Testable Assessments	% Correctly Classified
West Germany	20	80
Japan	20	80
India	18	83.3
China	15	86.7
Sweden	13	84.6
Canada	12	66.7
Israel	11	81.8
Italy	10	70
Netherlands	7	57.1
Belgium	6	50
NATO/European Grouping	6	83.3
Australia	5	60
Egypt	5	100
Switzerland	5	100
France	4	50
Norway	4	100
South Africa	3	33.3
Czechoslovakia	3	100
East Germany	3	100
Pakistan	2	100
Argentina	2	100
Brazil	2	100
Portugal	2	100
Romania	2	100
Spain	2	100
Taiwan	2	100
Albania	1	100
Bulgaria	1	100
Denmark	1	100
Hungary	1	100
Indonesia	1	100
Philippines	1	0
Poland	1	100
South Korea	1	0
Yugoslavia	1	100

Accounting for Characteristics and Composition of Countries

One natural concern is that the forecasts improved over time simply because the pool of countries being assessed shifted, or because the characteristics of these countries shifted to make them easier to assess. In order to account for this, I estimate multivariate linear probability models, including a dummy variable for post-1958 NIEs and employing as the dependent variable the basic measure of discrimination discussed in the manuscript: i.e. an assessment is correctly classified (coded as 1) if one of two conditions hold: greater than 50% odds are attached and the event does in fact occur in the time frame specified, or less than 50% odds are attached and the event does not occur in the specified time frame. Conversely, assessments are coded as incorrectly classified (coded 0) if less then 50% odds are attached and the event does occur in the time frame, or greater than 50% odds are attached and the event does not occur in the time frame. In separate models, I use dependent variables that are coded 1 for false positives and negatives.

I use linear probability models for their ease of interpretation; all results are similar using logit or probit models. As control variables, I include two binary measures of regime type,[1] dummy variables indicating whether the target country is a formal U.S. ally or adversary,[2] two measures of the target country's security environment,[3] dummy variables for whether the target country has a nuclear energy program or a peaceful nuclear cooperation agreement with the United States,[4] and a dummy variable measuring whether the assessment is judging current or

[1] These include whether the target country is a strong democracy (polity score greater than 7) or strong autocracy (polity score less than -7). Data is from Marshall, Gurr, and Jaggers 2018.
[2] Alliance data is from Leeds et al 2002. I code as U.S. adversaries Communist China, Warsaw Pact states, Indonesia in 1966, and Egypt from 1963-1966.
[3] These include a binary measure of involvement in an interstate rivalry and a moving average of the number of militarized disputes in the past five years. Data from Singh and Way 2004.
[4] I code a nuclear energy program as present if the target country has a power reactor operating or under construction. Data from World Nuclear Association 2018. Data on U.S. nuclear cooperation agreements is from Fuhrmann 2012.

future intentions. Collectively, these variables proxy for alternative explanations drawn from the existing literature: democratic states should be easier to assess; assessments of adversaries and countries with greater capabilities (nuclear energy program) or in threatening security environment should be prone to overestimates; and countries allied with the United States (or that have nuclear cooperation agreements with the U.S.) should be easier to assess due to a history of close interaction. Finally, for obvious reasons, assessments of current intentions should be easier than projections about future intent.

Table A2: Regression Models, 1957-1966

	Basic Discrimination	False Positives	False Negatives	Basic Discrimination, Fixed Effects
1960-66 Dummy	0.396	-0.383	-0.014	0.434
	(0.077)***	(0.076)***	(0.041)	(0.092)***
Adversary	0.154	0.091	-0.245	-1.722
	(0.112)	(0.112)	(0.094)**	(0.975)*
Ally	-0.031	0.093	-0.062	-1.070
	(0.063)	(0.048)*	(0.055)	(0.528)**
Democracy	-0.011	0.146	-0.135	-0.639
	(0.112)	(0.112)	(0.098)	(0.420)
Autocracy	0.109	-0.028	-0.081	0.005
	(0.058)*	(0.061)	(0.048)*	(0.363)
Recent Disputes	0.004	-0.024	0.020	0.046
	(0.009)	(0.013)*	(0.012)	(0.045)
Rivalry	0.073	-0.033	-0.039	
	(0.043)*	(0.066)	(0.053)	
U.S. NCA	0.033	0.007	-0.040	-0.093
	(0.051)	(0.040)	(0.034)	(0.198)
Nuclear Energy	0.095	-0.054	-0.041	0.098
	(0.052)*	(0.059)	(0.045)	(0.117)
Future	-0.086	0.055	0.031	-0.070
	(0.051)	(0.031)*	(0.042)	(0.085)
Constant	0.424	0.309	0.267	1.800
	(0.180)**	(0.176)*	(0.128)**	(0.726)**
R^2	0.24	0.31	0.08	0.37
N	187	187	187	187

* $p<0.1$; ** $p<0.05$; *** $p<0.01$ Blank entries indicate variables omitted due to collinearity

The results confirm the finding in the manuscript: assessments were substantially better in the 1960s, and much less likely to be false positives. This holds accounting for country fixed effects, suggesting a changing composition of countries is not responsible for the results. Indeed, there are seven countries that the NIEs errantly ascribed proliferation intent to at least twice in the 1950s and then assessed accurately at significantly higher rates in the 1960s: Belgium, Canada, Italy, Japan, the Netherlands, Sweden, and West Germany. This does not establish a causal relationship, but it shows that the temporal change cannot be accounted for by the composition of countries being analyzed or measurable characteristics like regime type.

Results Substituting Bleek Codings

In order to ensure the results are not driven by using Singh and Way's coding of nuclear pursuit, I replicated the analysis using Bleek's codings and found similar patterns: an overall rate of 75.3% assessments correctly classified, substantial improvement on all metrics in the 1960s, and a lower rate of false positives in particular.

Table A3

	Full Sample	1957-58	1960-66
Discrimination, % correctly classified	75.3%	48.9%	82.7%
Brier Score	.198	.381	.147
Calibration Index	.039	.188	.023

Figure A1

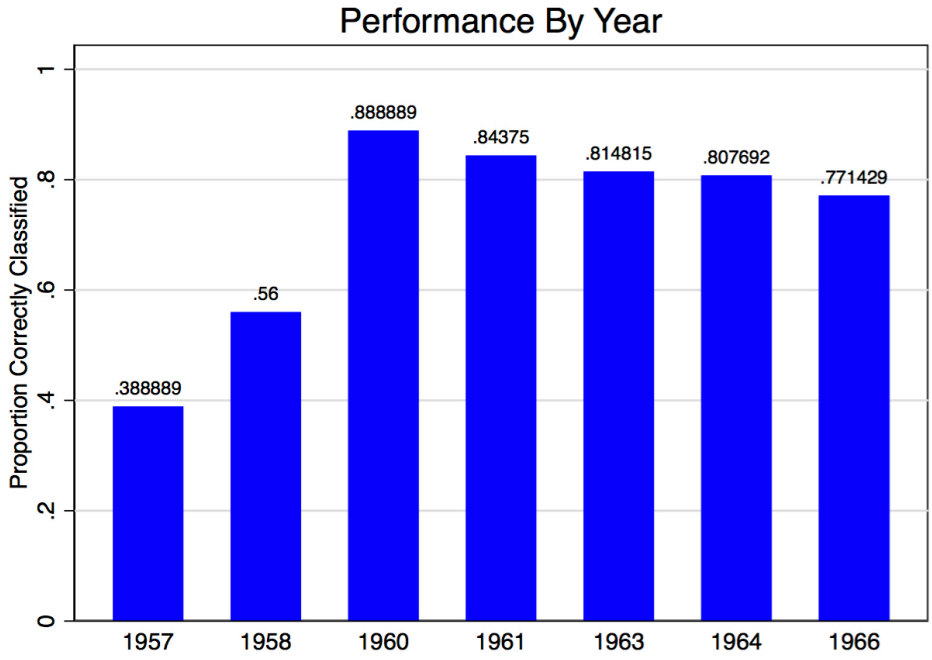
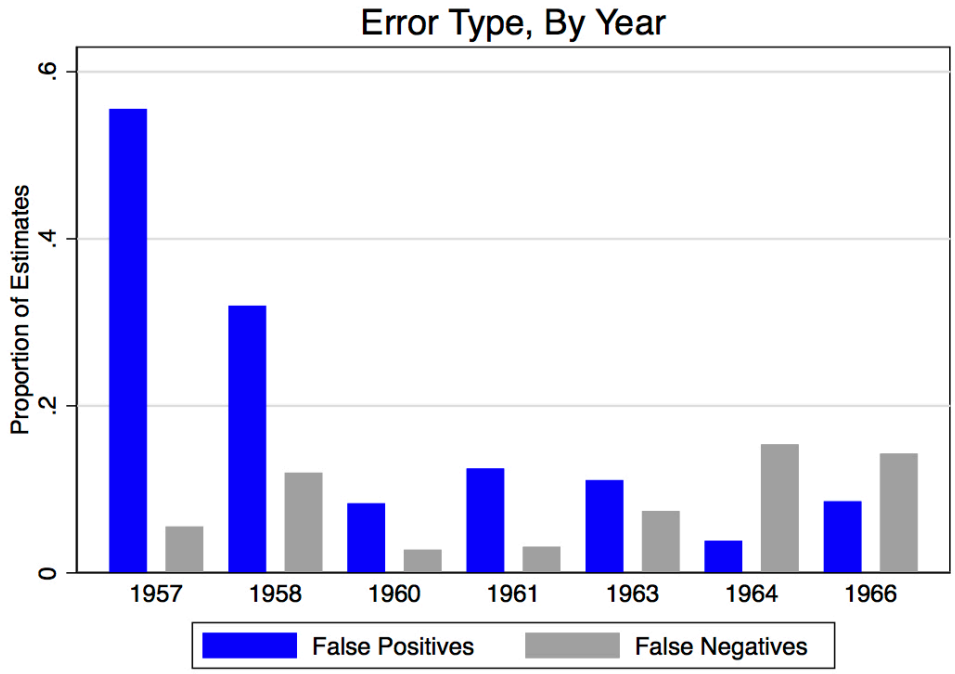


Figure A2


Error Type, By Year
Proportion of Estimates
0
.2
.4
.6
1957
1958
1960
1961
1963
1964
1966
False Positives
False Negatives

Results Only Using Outcome Assessments

To assess whether process assessments are affecting the results, I replicated the analysis while restricting the dataset only to outcome assessments and found similar results: an overall rate of 83.3% assessments correctly classified, substantial improvement on all metrics in the 1960s, and a significantly lower rate of false positives in particular.

Table A4

	Full Sample	1957-58	1960-66
Discrimination, % correctly classified	83.3%	54.5%	90.7%
Brier Score	.141	.364	.083
Calibration Index	.032	.188	.007

Figure A3

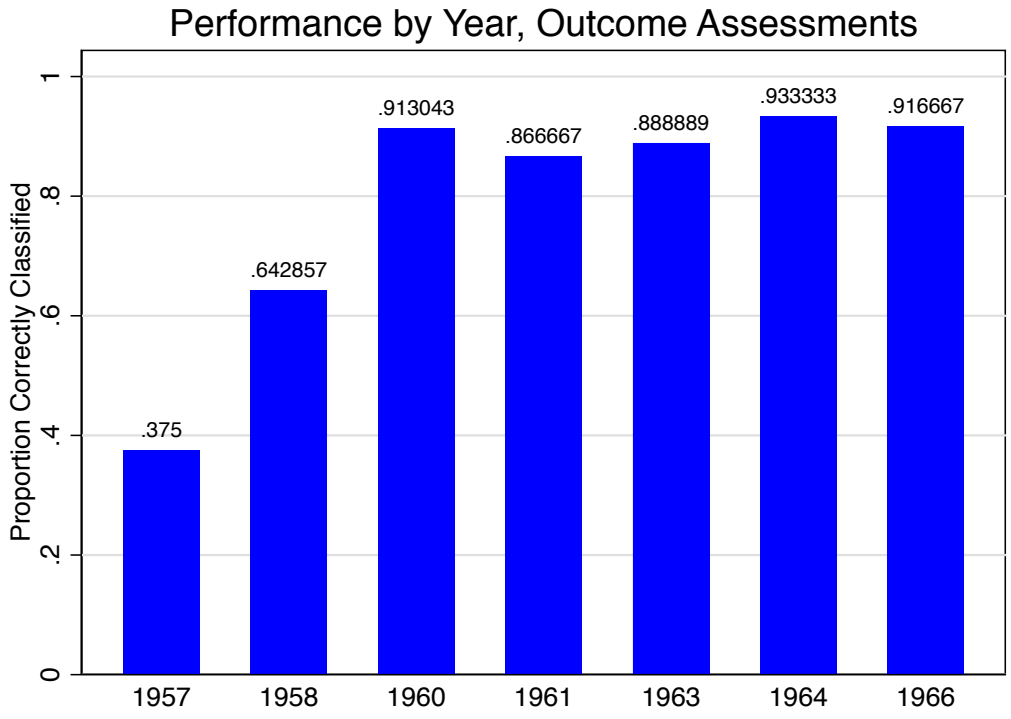
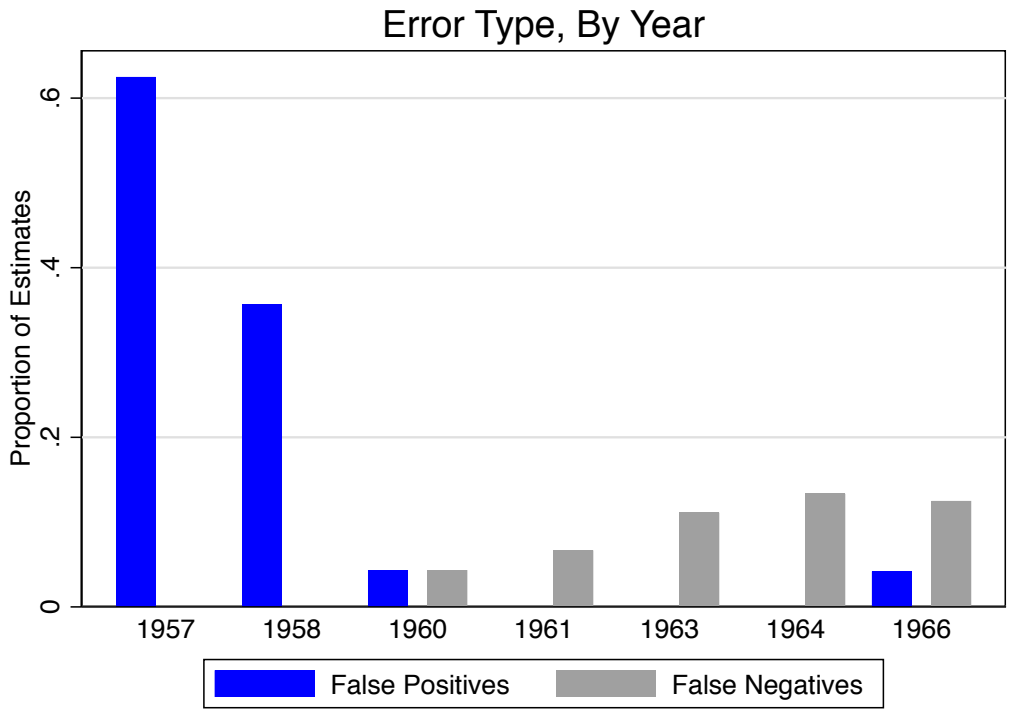


Figure A4


Error Type, By Year
Proportion of Estimates
0
.2
.4
.6
1957
1958
1960
1961
1963
1964
1966
False Positives
False Negatives

Robustness to Including 1974 NIE

Another potential concern is that the results only apply to a narrow slice of time. Perhaps assessments reverted to being inaccurate in the 1970s, as countries reacted to the presence of the NPT by acting more covertly. To test for this, I coded a similar 1974 NIE and found that the results remained much closer to the performance in the 1960s than the 1950s: 89% correctly classified using the basic measure of discrimination, a Brier score of .118, and calibration index of .026.[5]

[5] I located two additional declassified proliferation estimates, from 1982 and 1985. However, these estimates shifted from using a ten year window as the default for predictions to a five year window, thus making it an apples to oranges comparison with the earlier NIEs.

Explaining Change over Time, Supplementary Analysis

Table A5: Brier Scores by Time Period and Country Type

	Brier Score for Intel Targets	**Score w/ Base Rate by Era**
1957-1958	.520	.243
1960-1966	.065	.230

Table A6: Regression Models Accounting for Change over Time

	Intel Targets	Non-Targets	Intel Targets, Fixed Effects	Non-Targets, Fixed Effects
1960-66 Dummy	0.700	0.143	0.728	0.068
	(0.057)***	(0.086)	(0.072)***	(0.054)
Ally	-0.164	-0.011		
	(0.033)***	(0.140)		
Democracy	-0.619	0.003	-0.684	
	(0.066)***	(0.165)	(0.033)***	
Recent Disputes	-0.007	-0.010	-0.029	-0.002
	(0.013)	(0.022)	(0.055)	(0.039)
Rivalry	0.238	0.094		
	(0.053)***	(0.140)		
Nuclear Energy	-0.108	0.079	-0.156	0.273
	(0.063)	(0.056)	(0.117)	(0.170)
Future	-0.109	-0.064	-0.106	-0.094
	(0.062)	(0.090)	(0.059)	(0.110)
Adversary		0.163		
		(0.219)		
Autocracy		0.072		-0.000
		(0.068)		(0.004)
U.S. NCA		0.001		-0.181
		(0.067)		(0.119)
Constant	1.055	0.650	1.139	0.899
	(0.078)***	(0.304)**	(0.079)***	(0.150)***
R^2	0.52	0.09	0.52	0.30
N	85	102	85	102

* $p<0.1$; ** $p<0.05$; *** $p<0.01$ Blank entries indicate variables omitted due to collinearity

Raw Data on Calibration

Table A7: Full Sample (1957-1966)

Forecast Probability	Actual Frequency	n
0	.083	36
.035	0	7
.070	0	1
.100	0	3
.200	.143	28
.250	.500	4
.450	0	1
.550	1	2
.600	0	1
.750	.620	50
.950	.813	16
1	.778	45

Table A8: Assessments from 1957-1958

Forecast Probability	Actual Frequency	n
0	0	1
.200	0	3
.250	0	1
.600	0	1
.750	.352	17
.950	.800	5
1	.429	14

Table A9: Assessments from 1960-1966

Forecast Probability	Actual Frequency	n
0	.086	35
.035	0	7
.070	0	1
.100	0	3
.200	.160	25
.250	.667	3
.450	0	1
.550	1	2
.750	.758	33
.950	.818	11
1	.935	31

Alternative Explanations for Improvement over Time

There are at least five alternative explanations that might account for the improvement in the accuracy of assessments: a greater high-level political focus on nonproliferation, the emergence of the nonproliferation regime, turnover in who is doing the assessments, changes in technology, and differences in the frequency of vague or specific assessments

With respect to the impact of high-level political focus, there is at least face validity to the notion that this contributed to improved accuracy rates. After all, accuracy was higher under the Kennedy and Johnson administrations—two presidential administrations that were relatively strong in their commitment to preventing proliferation—and lower under Eisenhower, who was notably lax on the issue.[6] In other words, it is possible that Kennedy and Johnson intervened to make proliferation a higher priority target for intelligence agencies, thus leading to increased resources and attention and subsequently higher accuracy. Indeed, in January 1961, shortly after Kennedy entered the White House, the Joint Atomic Energy Intelligence Committee completed an intelligence post-mortem on the failure to detect Israel's nuclear weapons program in a timely fashion.[7] The report recommended a greatly increased attention to the problem of proliferation intelligence worldwide, including more efforts at collection, greater information sharing within the U.S. intelligence community, and "a concerted effort…to obtain full reporting on the political factors that would identify the motivations or intentions of potential 'Nth' countries to pursue a nuclear weapons capability."[8] Yet there is a problem with attributing the increased accuracy to the newfound political priority given to proliferation: as Figure 1 in the main manuscript makes

[6] Miller 2018.
[7] Cohen 1998, 4. On U.S. intelligence toward Israel's nuclear program, see Long and Shifrinson 2019.
[8] U.S. Joint Atomic Energy Intelligence Committee 1961.

clear, the improvement had already occurred by 1960, even if it was subsequently supported by policy changes put in place under Kennedy and Johnson.

A second possibility is that assessments became more accurate over time due to the emergence of the nonproliferation regime, which forced states to more clearly state their position on nuclear weapons and might have made clear that the overall trend was against widespread proliferation. This problem is partly addressed by design; as noted above, the main analysis only looks at assessments of countries before the NPT was concluded in 1968. Moreover, the most significant development in the nonproliferation regime prior to the NPT came in 1963 with the Limited Test Ban Treaty, but this was three years after the improvement in accuracy began. A qualitative reading of the 1964 and 1966 NIEs suggest that the treaty was not a significant factor playing into assessments; indeed, the 1964 NIE explicitly noted, "The 1963 partial nuclear test ban treaty, which permits only underground tests, does not pose a significant technical problem for a small-scale weapons program."[9] Further, accuracy significantly improved in the 1960s both for assessments predicting proliferation and nonproliferation, as Table A9 below shows. This suggests that improvement is not solely driven by the IC recognizing in the 1960s that proliferation would be much more limited than previously thought.

Table A10: Accuracy of Nonproliferation vs. Proliferation Assessments, by Time Period

	Nonproliferation Predictions	Proliferation Predictions
1957-1958	63.6% (n=11)	43.8% (n=32)
1960-1966	87.6% (n=105)	84.3% (n=51)
p-value from chi-square test	.032	<.00

[9] NSA, EBB 401, doc. 3.

The quantitative findings also cannot be explained by bilateral safeguards required in U.S. transfers of nuclear technology: we still see improvement over time even controlling for whether a country had signed a nuclear cooperation agreement with the United States. In fact, the countries where intelligence assessments improved had already signed such agreements with the United States by 1957, suggesting this cannot explain the sharp improvement in 1960.

Third, what about turnover within the intelligence community itself? It's possible that more skilled or talented analysts became involved over time (though the opposite is of course possible too). It is obviously not possible to gather comprehensive data on all intelligence officers involved in the analysis and collection of information relevant to the NIEs, but there is data available from much of the time period under study on membership of the Board of National Estimates (BNE), the group headed by Sherman Kent that supervised and was ultimately responsible for producing NIEs.[10] Table 5 below shows the attrition in Board members from 1957—when the first NIE under analysis was produced—to 1963, when the available data on Board membership ends. The data show that there was significant turnover between 1957 and 1958—when there was a corresponding increase in accuracy rate—but no similarly noticeable turnover from 1958 to 1960, when there was another significant increase in accuracy rate. This evidence is by no means dispositive, especially since it may be that the lower-level analysts are more important for estimate quality than the higher-level officials on the BNE. However, this is the best that can currently be done given data constraints.

[10] See Kent 1994, appendix.

Table A11: Turnover in Board of National Estimates Membership, 1957-1963

Year	Basic discrimination (% correctly classified)	Fraction of Members Who Served in 1957
1957	38.9	13/13
1958	56.0	8/13
1960	91.7	8/13
1961	81.3	6/13
1963	85.2	5/12

Fourth, what if assessments improved not because of learning but rather because of technological advances that increased the amount or quality of information available to the assessors? This has some plausibility on its face, as the U.S. government introduced the Corona satellite system in mid-1959, right before U.S. assessments began to improve. Corona offered the intelligence community a less risky means of obtaining overhead photography of foreign countries, as well as the ability to photograph much wider areas.[11] However, based on what we know about which foreign nuclear programs were targeted by Corona, this does not seem to account for the improving accuracy of assessments we observe. Instead of targeting U.S. allies in Europe (plus Japan) where estimates improved after 1958, Corona reconnaissance focused on a different set of nuclear programs, namely China, India, Israel, Taiwan, and North Korea.[12] Only two of these countries were assessed both before and after 1958 and one of them saw improved accuracy (China) and one saw reduced accuracy (Israel). Indeed, while we would expect photographic intelligence to improve assessments of *capabilities*, it is far less clear if we should expect this effect on assessments of intentions, particularly given the dual-use nature of most nuclear installations. For example, U.S. analysts found Corona photographs of India's nuclear facilities to be of quite limited use in this regard. As Richelson puts it, "Whether India would,

[11] See Day, Lodgson, and Latell 1998.
[12] See Richelson 2007. Corona satellites did gather intelligence on the French nuclear program, but this began after France had already acquired nuclear weapons.

sometime in the future, decide to join the nuclear club was a mystery to analysts for years—a question that no one, not even the Indian government itself, could answer. Collection systems such as Corona and Gambit were of no use in trying to unravel such mysteries."[13]

A final possibility is that the improved accuracy rate is a statistical artifact caused by the coding rules, which excluded assessments whose accuracy could not be coded due to vague language like "could," "might," or "possibly." It could be that the increasing accuracy rate is caused by the fact that intelligence assessors were taking more risks early on and making clearer predictions—even in cases where confidence was relatively low—and taking fewer risks later and only making clear predictions when confidence was high. If this was the case, we would expect to see a higher proportion of vague assessments in later years, when the accuracy rate was higher. Table 7 below shows that this is not the case: although about 20% of all assessments of proliferation intentions were too vague to be included in the analysis, there is no clear trend of a greater proportion of vague assessments being made over time. In fact, the rate of vague assessments was somewhat higher in the 1950s compared to the 1960s.

Table A12: Frequency of Vague Assessments Over Time

Year	**# of Testable Assessments**	**# of Vague Assessments**	**% Vague out of Total #**
1957	18	6	25.0
1958	25	8	24.2
1960	36	10	21.7
1961	32	7	17.9
1963	27	9	25.0
1964	26	9	25.7
1966	35	6	14.6
TOTAL	199	54	21.3

[13] Ibid., 228.

Relatedly, what if earlier assessments were simply more specific in terms of time frame, thereby being both more useful to policymakers but also more likely to be incorrect? To test this, I coded whether each assessment attached a more specific time frame to its estimate than the default ten-year window which the NIEs were explicitly guided to assess. Controlling for this specificity variable in a regression (Table A11 below) slightly attenuates but does not change the finding that assessments became significantly more accurate over time, especially for countries identified as targets in 1958.

Table A13: Accounting for Specificity of Assessments

	Full Sample	Intel Targets	Non-Targets
1960-66 Dummy	0.345 (0.077)***	0.627 (0.072)***	0.141 (0.081)*
Adversary	0.158 (0.111)		0.168 (0.221)
Ally	-0.074 (0.072)	-0.221 (0.035)***	-0.011 (0.141)
Democracy	0.010 (0.110)	-0.540 (0.084)***	0.004 (0.166)
Autocracy	0.128 (0.060)**		0.075 (0.068)
Recent Disputes	0.001 (0.010)	-0.001 (0.012)	-0.006 (0.022)
Rivalry	0.075 (0.048)	0.232 (0.049)***	0.090 (0.143)
U.S. NCA	0.055 (0.057)		0.007 (0.071)
Nuclear Energy	0.110 (0.055)*	-0.060 (0.067)	0.076 (0.057)
Future	-0.071 (0.054)	-0.085 (0.063)	-0.047 (0.085)
Specific	-0.336 (0.113)***	-0.222 (0.098)*	-0.328 (0.376)
Constant	0.452 (0.183)**	1.033 (0.091)***	0.629 (0.309)*
R^2	0.27	0.53	0.10
N	187	85	102

* $p<0.1$; ** $p<0.05$; *** $p<0.01$ Blank entries indicate variables omitted due to collinearity

Assessing Quality of Reasoning

To increase confidence in the findings of the Sweden case study, I also examined the quality of reasoning deployed in assessments of West Germany and Japan in the 1960s—countries with a large number of accurate assessments whose accuracy improved over time. Overall, comparing the assessments to subsequent historical accounts suggests the intelligence analysts were mostly right for the right reasons. I briefly describe the results of this exercise below.

West Germany

After early estimates in the 1950s errantly predicted West Germany would seek its own nuclear weapons or an independent European nuclear deterrent, assessments notably improved in the 1960s. Estimates in 1960 and 1961 correctly noted that German concerns about U.S. alliance credibility would lead them to consider the nuclear option and that worries about the Soviet response and domestic political obstacles would inhibit them from going down the nuclear path. These estimates also correctly reasoned that West Germany would prefer a multilateral nuclear approach under U.S. auspices (like the ill-fated MLF) to a unilateral capability. However, the reasoning wasn't perfect, seeming to over-emphasize the importance of treaty restrictions (particularly the 1954 Paris Accords) and not mentioning the importance of U.S. opposition. The 1963-1966 estimates were similar but correctly added in West German concerns about the reaction of its allies (the United States in particular) as an important inhibiting factor. Like earlier estimates, these probably over-emphasized the importance of treaty commitments and technical obstacles.

Sources:

Gerzhoy, Gene. 2015. Alliance Coercion and Nuclear Restraint: How the United States Thwarted West Germany's Nuclear Ambitions. *International Security* 39 (4): 92-129.

Lutsch. Andreas. 2015. The Persistent Legacy: Germany's Place in the Nuclear Order. *Nuclear Proliferation International History Project,* Working Paper #5.

Monteiro, Nuno and Alexandre Debs. 2014. The Strategic Logic of Nuclear Proliferation. *International Security* 39 (2): 7-51.

Japan

After overemphasizing Japanese interest in nuclear weapons in the 1950s, starting in 1960 the assessments correctly keyed in on several inhibiting factors. This NIE accurately noted the importance of cost in weighing in Japanese decision-making and accurately forecast that China acquiring nuclear weapons would lead Japan to more seriously consider acquiring its own nuclear arsenal. It over-emphasized, however, the importance of domestic opposition to nuclear weapons, a consistent pattern in the 1960s estimates that conflicts with much of the subsequent literature on what drove Japanese decision-making. Estimates from 1961 and 1963 were similar but also rightly observed that there was strong support for the alliance with the United States as an alternative to an independent arsenal, that there were major risks of a nuclear capability given Japan's concentrated population (and vulnerability to nuclear attack), and that Japan could follow a more attractive middle path of keeping the nuclear option open technologically without starting a weapons program. The 1964 and 1966 NIEs deployed similar reasoning as the prior estimates, with the 1966 estimate correctly adding in concerns about U.S. opposition as a constraining factor.

Sources:

Kase, Yuri. 2001. The Costs and Benefits of Japan's Nuclearization: An Insight into the 1968/70 Internal Report. *Nonproliferation Review* 8 (2): 55-68.

Hughes, Lleweyln. 2007. Why Japan Will Not Go Nuclear (Yet): International and Domestic Constraints on the Nuclearization of Japan. *International Security* 31 (4): 67-96.

Hymans, Jacques. 2011. Veto Players, Nuclear Energy, and Nonproliferation: Domestic Institutional Barriers to a Japanese Bomb. *International Security* 36 (2): 154-189.

Rublee, Maria Rost. 2008. Taking Stock of the Nuclear Nonproliferation Regime: Using Social Psychology to Understand Regime Effectiveness. *International Studies Review* 10 (3): 420-450.