Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Professionals

Jeffrey A. Friedman, Jennifer S. Lerner, and Richard Zeckhauser

Abstract National security is one of many fields where experts make vague probability assessments when evaluating high-stakes decisions. This practice has always been controversial, and it is often justified on the grounds that making probability assessments too precise could bias analysts or decision makers. Yet these claims have rarely been submitted to rigorous testing. In this paper, we specify behavioral concerns about probabilistic precision into falsifiable hypotheses which we evaluate through survey experiments involving national security professionals. Contrary to conventional wisdom, we find that decision makers responding to quantitative probability assessments are less willing to support risky actions and more receptive to gathering additional information. Yet we also find that when respondents estimate probabilities themselves, quantification magnifies overconfidence, particularly among low-performing assessors. These results hone wide-ranging concerns about probabilistic precision into a specific and previously undocumented bias that training may be able to correct.

Although uncertainty surrounds nearly every national security decision, national security officials are often reluctant to assess this uncertainty directly.[1] For example, when General Stanley McChrystal recommended deploying 40,000 additional soldiers to Afghanistan in 2009, he explained to President Obama that this would "improve effectiveness" and that it offered "the best prospect for success in

For comments on drafts and research design, we thank Mark Bucknam, Erik Dahl, Ryan Enos, Joshua Kertzer, Paul Novosad, Brendan Nyhan, Bryan Pendleton, Chris Robert, Kathryn Schwartz, and Peter Scoblic. Joowon Kim and Max Yakubovich provided outstanding research assistance. Previous versions of this paper were presented to Dartmouth's Government Department, Harvard's Workshop on Political Psychology, MIT's Political Science Department, and the 2015 annual meetings of the International Studies Association and the Midwest Political Science Association. We are particularly grateful to the 407 national security officials who volunteered their time to participate in this research. This research was supported by the United States Department of Homeland Security (DHS) through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) at the University of Southern California (USC) under award number 2010-ST-061-RE0001. Any opinions, findings, conclusions, or recommendations in this document are those of the authors and do not necessarily reflect views of DHS, the University of Southern California, or CREATE.

1. In the decision sciences, *uncertainty* refers to situations where probabilities cannot be estimated precisely. This contrasts to situations of "risk," such as playing roulette, where relevant probabilities are known.

International Organization, 2017, page 1 of 24

doi:10.1017/S0020818317000352

this important mission."[2] Yet even if some action offers the best chances of success, this does not imply that those chances are worth accepting. The key question was not whether the Afghan Surge would raise the chances of achieving a favorable outcome, but whether this increase was large *enough* to justify the policy's expected costs, and McChrystal's report did not explicitly address this issue.

Vague probability assessments are both common and deliberate in national security decision making. Figure 1, for example, shows three sets of guidelines instructing US intelligence analysts to describe uncertainty using qualitative language.[3] US military doctrine instructs planners to identify actions that maximize the chances of success, but not necessarily to identify what those chances are.[4] The Department of Homeland Security divides terrorist threats into tiers ("elevated," "intermediate," and "imminent") that lack clear probabilistic equivalents.[5] Similar debates about communicating uncertainty surround the conduct of national security analysis in Britain, Canada, and Israel, to name just a few examples.[6] And when scholars, practitioners, and pundits debate national security decisions in the broader marketplace of ideas, they often neglect to describe key probabilistic assumptions: recommending policies, for example, without describing the chances that these policies will succeed.[7]

Many observers find this behavior problematic.[8] After all, probability is an inherently quantitative concept, representing values between 0 and 100 percent. Analysts always have a coherent conceptual basis for quantifying probability estimates, no matter how subjective those estimates might be.[9] If analysts conveyed probability assessments using numbers, then these assessments might not always be accurate, but at least they would be clear. There would be no need to worry about what it means to speak about the "best prospect for success" in Afghanistan or to wonder what an "intermediate" threat of terrorism implies.

Yet many scholars and practitioners worry that even if making probability assessments more precise would support rigorous decision making in principle, this could lead to harmful consequences in practice. One prominent concern is that analytic precision creates illusions of rigor, such that quantifying probability assessments would cause decision makers to see these estimates as being more scientific than they really are. A second prominent concern is that many people find quantitative reasoning counterintuitive, such that quantifying subjective probability assessments would be like speaking in a second language, inducing otherwise avoidable errors.

2. McChrystal 2009.

3. The Defense Intelligence Agency memorandum further explains that "*DIA does not condone the use of probability percentages in its products to portray likelihood*" (emphasis in original, Tradecraft Note 01-15: Expressing Analytic Certainty, 5 January 2015).

4. See US Army 2009, paragraphs 2-19, B-173; US Army 1997, paragraphs 5-24.

5. This replaced a previous system of color-coded threat warnings that may have been even more problematic. See Shapiro and Cohen 2007.

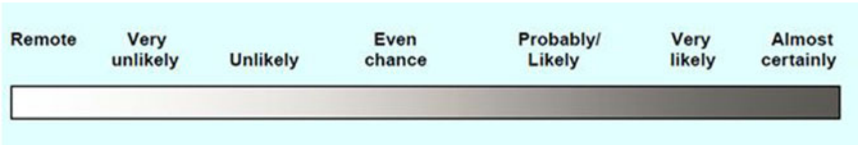
6. See Barnes 2015; Dhami 2013; and Lanir and Kahneman 2006.

7. Tetlock and Gardner 2015.

8. Controversy over this subject dates back to Kent 1964.

9. Savage 1954.

A *Guidelines "Explaining Estimative Language" in National Intelligence Estimates (2007)*



B *US Director of National Intelligence, Intelligence Community Directive 203, "Analytic Standards" (2015)*

(a) For expressions of likelihood or probability, an analytic product must use one of the following sets of terms:

almost no chance	very unlikely	unlikely	roughly even chance	likely	very likely	almost certain(ly)
remote	highly improbable	improbable (improbably)	roughly even odds	probable (probably)	highly probable	nearly certain
01-05%	05-20%	20-45%	45-55%	55-80%	80-95%	95-99%

C *Defense Intelligence Agency memorandum on "Expressing Analytic Certainty" (2015)*

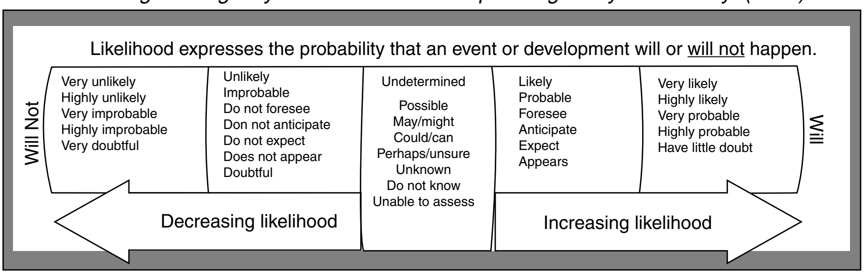


FIGURE 1. *Three guidelines for expressing probability in intelligence*

Both of these arguments suggest that efforts to make probability assessments clearer and more rigorous could actually impair the way that analysts and decision makers assess uncertainty. And this concern has major implications for theory and practice. Following the broader trend of exploring behavioral decision making throughout the social sciences, international relations scholars have demonstrated that national security officials are subject to a broad range of heuristics and biases when making high-stakes decisions.[10] This scholarship is important because it suggests that if analysts and decision makers scrutinized their judgments more carefully, it would be possible to improve their behavior.[11] But the "illusions of rigor" and "numbers as a second language" arguments warn that attempts to analyze

10. See Hafner-Burton et al. 2017 and Levy 2013.
11. Following a standard distinction in the decision sciences, we use the term *decision makers* to refer to individuals who are interpreting probability assessments and *analysts* to refer to individuals who are making probability assessments.

probabilistic reasoning explicitly can backfire, activating a different set of psychological distortions that could create major problems of their own. Given how the challenge of assessing uncertainty surrounds any intelligence report, military plan, or foreign policy debate, these issues have broad relevance for efforts to understand and improve national security decision making. Yet we are not aware of any existing research that tests these arguments directly.

This study hones the illusions-of-rigor and numbers-as-a-second-language arguments into six testable hypotheses, and then tests them through preregistered survey experiments administered to 407 national security professionals and to 3,017 respondents from Amazon Mechanical Turk. Our results roundly contradict the illusions-of-rigor argument. Contrary to conventional wisdom, we find that quantifying probability assessments makes decision makers less willing to support proposed actions and more receptive to gathering additional information. Yet our results also support a particular version of the numbers-as-a-second-language argument. We find that when respondents estimate probabilities themselves, quantification magnifies their general tendency to express judgments with unjustified certitude, especially among low-performing assessors. While these results hardly close debates about the costs and benefits of vague probability assessments, they hone a broad list of behavioral concerns into a specific and previously undocumented bias that training may be able to correct.

Of course, national security is just one of many fields where it is worth debating proper methods for assessing and communicating uncertainty. Medicine, law, regulation, and climate science are four additional domains that feature similar debates about the value and limits of precision when assessing probability.[12] In the broadest sense, this study hones wide-ranging skepticism about the behavioral consequences of probabilistic precision into falsifiable claims, develops an original empirical methodology for evaluating those claims, and shows how several common objections to explicit probability assessments do not hold up to direct tests in either elite or non-elite samples.

Uncertainty, Complexity, and Precision in National Security Decision Making

One of the central problems with assessing uncertainty in national security decision making is that the most important judgments also tend to be the most subjective and contentious.[13] Mark Lowenthal thus warns, in arguably the most important textbook on intelligence studies, that numeric precision conveys "a degree of precision that does not exist" when assessing probability.[14] The 2007 National Intelligence Estimate, *Prospects for Iraq's Stability*, similarly justifies the use of vague probability estimates by explaining to readers that "assigning precise numerical ratings to

12. See Budescu et al. 2014; Gigerenzer 2002; Sunstein 2014; and Tillers and Gottfried 2006.
13. See Betts 2006; Beyerchen 1992/93; Fingar 2011; Jervis 1997; and Rovner 2011.
14. Lowenthal 2006, 129.

such judgments would imply more rigor than we intend."[15] These statements imply not just that numeric probability assessments represent arbitrary detail, but that this detail could reduce the quality of national security decision making. This is what we call the "illusions-of-rigor" thesis.

The basic assumption driving the illusions-of-rigor thesis is that greater degrees of analytic precision could cause decision makers to infer that they possess a stronger evidentiary basis for evaluating choices under uncertainty. And there are at least three distinct possibilities for how exaggerating the strength of available evidence could impair the quality of high-stakes decision making. Each of these hypotheses represents a different idea of what it means to say that probabilistic precision implies "stronger" judgments.

First, quantifying probability assessments could alleviate decision makers' concerns about placing personnel, resources, and national interests at risk based on incomplete information. In a criminal trial, for example, juries are not supposed to convict defendants simply because they think that there is a high probability that the defendant is guilty. In principle, jurors should also believe that there is reliable evidence supporting this view. For similar reasons, national security decision makers may be reluctant to take proposed actions if they question the reliability of their assumptions. Prior to the invasion of Iraq, for example, public officials created the impression that the United States possessed reliable intelligence indicating that Saddam Hussein was developing weapons of mass destruction (WMDs), whereas those judgments actually depended on circumstantial evidence and questionable informants. If the speculative nature of these judgments had been clearer, then even if senior officials in Congress or the executive branch still believed it was likely that Iraq was pursuing WMDs, they might have found it harder to justify pursuing regime change.[16]

By a similar logic, if quantitative probability assessments appear to provide a reliable basis for making decisions, then analytic precision could increase decision makers' willingness to support proposed actions. This argument relates to the well-known phenomenon of "ambiguity aversion," in which decision makers are more willing to bet on probabilities that are known as opposed to those that are ambiguous.[17] But while ambiguity aversion is generally thought to be irrational, one can argue that this behavior is appropriate in the context of national security decision making, either as a result of ethical concerns surrounding the use of force, or due to behavioral concerns that analytic precision could exacerbate national security decision makers' natural tendencies towards overconfidence. Either way, one could worry that quantifying probability assessments would increase national security decision

15. National Intelligence Council National Intelligence Council, Iran: Nuclear Intentions and Capabilities, press release November 2007, accessed 21 August 2017 from <https://www.dni.gov/files/documents/Newsroom/Press%20Releases/2007%20Press%20Releases/20071203_release.pdf>.

16. On the Iraq WMD Estimate's use of evidence, see Betts 2006 and Jervis 2010.

17. Ellsberg 1961. In other words ambiguity-averse decision makers have lower probabilistic thresholds for taking gambles based on probabilities that are more precise.

makers' willingness to support proposed actions, and this is our first hypothesis for evaluating the illusions of rigor thesis.

H1a: Quantifying probability assessments increases support for proposed actions.

A second way of specifying the problems that illusions of rigor might cause is to say that probabilistic precision does not necessarily bias decision makers towards taking proposed actions, but rather that it amplifies the weight that decision makers assign to analysts' judgments. In this view, probability assessments are just one of many cues that decision makers use to make choices under uncertainty. When analysts express probabilities more precisely, decision makers may believe that those judgments are more credible.

In contrast to our first hypothesis, this argument does not imply that probabilistic precision biases decision makers towards or away from taking proposed actions on the whole. If analysts quantified a seemingly favorable probability assessment, such as a high chance that a hostage rescue mission will succeed, this could make decision makers more willing to support the proposal. By contrast, if analysts quantified a seemingly unfavorable probability assessment, such as a high chance that a drone strike will cause collateral damage, this could make decision makers *less* willing to support proposed actions. If quantifying probability assessments thus causes decision makers to assign undue weight to analysts' subjective judgments, this would represent a second way in which illusions of rigor could impair the quality of national security decision making.

H1b: Quantifying probability assessments amplifies support or disapproval for proposed actions.

A third way that illusions of rigor could impair decision making is by reducing decision makers' willingness to gather additional information when evaluating proposed actions. Indeed, this is perhaps the most straightforward implication of the idea that precise probability estimates seem more reliable than they really are. When dealing with uncertainty, decision makers frequently confront tradeoffs between acting immediately versus conducting additional analysis. Because conducting additional analysis carries costs—both the direct costs of gathering more information and the opportunity costs of delay—rational decision makers must consider the potential benefits that gathering this additional information might bring.

This third way of interpreting the illusions-of-rigor argument thus has less to do with biasing decision makers' levels of support for proposed actions. Rather, this concern pertains to how decision makers choose to *time* those proposed actions, and how they choose to structure processes for gathering and processing information.[18] If quantifying probability estimates leads decision makers to believe that these judgments are more reliable than they really are, then this could also cause

18. On information acquisition and foreign policy decision making, see Friedman and Zeckhauser 2015; Heuer 1999; and Mintz and Geva 1997.

decision makers to undervalue the benefits of delaying high-stakes choices. By extension, expressing probability assessments more precisely could encourage a potentially harmful rush to judgment.

H1c: Quantifying probability assessments reduces willingness to gather additional information before making decisions.

Thus while many observers argue that making probability assessments more explicit would encourage decision makers to consider the uncertainty surrounding their choices more carefully, Hypotheses 1a–1c suggest how attempts to clarify probability estimates could unintentionally backfire, warping decision makers' reactions to uncertainty instead of clarifying critical issues. While these arguments raise plausible behavioral concerns that are relevant to nearly any national security debate, there is currently little systematic empirical evidence supporting or refuting them. To our knowledge, no national security scholars or practitioners have submitted these claims to direct testing.

Numbers as a Second Language

Even if national security decision makers respond rationally to numeric probabilities, quantifying assessments of uncertainty could still prove detrimental if this degraded the content of the information that analysts provide. Some scholars argue that analysts naturally think about uncertainty qualitatively,[19] and many national security analysts are said to be especially uncomfortable expressing subjective judgments using numbers.[20] This perspective implies that quantifying probability assessments is like expressing complex ideas in a second language, conveying information in a format that induces avoidable errors in judgment.

H2a: Quantifying probability assessments reduces judgmental accuracy.

As with the illusions-of-rigor thesis, there are multiple ways in which quantifying subjective judgments could impair national security analysis, and these mechanisms have important practical implications. For example, if translating beliefs about uncertainty into numerical form simply adds random noise to the content of analysts' probability assessments, then this problem would be difficult to correct, either by training analysts to recalibrate their judgments or by explaining to decision makers how they might debias the judgments they receive. But if quantifying subjective judgments generates predictable errors, then it becomes more plausible to think that these errors could be identified and corrected. In particular, there are plausible reasons to believe that quantifying subjective probability assessments could systematically influence the degrees of certainty that analysts attach to their judgments.

19. Wallsten 1990; and Zimmer 1984.
20. Johnston 2005; and Kent 1964.

The first of these possibilities is that quantifying probability assessments causes analysts to "hedge" their bets, assigning too little certainty to their judgments. The most plausible reason this would be the case relates to concerns about accountability. Several scholars have argued that analysts prefer to make vague probability assessments because such vagueness allows them to redefine their judgments after the fact in a manner that deflects criticism, a practice known as "elastic redefinition."[21] Because explicit probability assessments foreclose opportunities for elastic redefinition, this could exacerbate analysts' concerns about receiving criticism when their judgments appear to be mistaken. To reduce this prospective criticism, analysts might choose to express their views with less certainty than these judgments deserve.

H2b: Quantifying probability estimates causes analysts to attach less certainty to their judgments.

There are also plausible reasons to expect the opposite effect, such that quantifying probability estimates would cause analysts to offer judgments with greater certitude. Since many probability assessors are naturally inclined to express uncertain judgments with excessive certainty,[22] one potential advantage of the kinds of guidelines shown in Figure 1 is that they impose natural anchors for calibrating judgments. Consider an analyst who believes that a statement has a high probability of being true, but still wishes to convey the presence of residual uncertainty. When expressing this judgment numerically, the analyst might anchor this judgment on certainty, and adjust it down to 90 or 95 percent. Using the guidelines in Figure 1, by contrast, an analyst wishing to signal residual uncertainty might instead select a term like "very likely," which, according to the Director of National Intelligence, covers probabilities as low as 80 percent. In this way, coarsening probability assessments could play a helpful role in mitigating analysts' tendencies to make judgments with unjustified certitude.

H2c: Quantifying probability estimates causes analysts to attach more certainty to their judgments.

Hypotheses 2a–2c thus stake a second set of important claims about how efforts to improve the clarity of national security analysis could perversely reduce the quality of the assessments that analysts provide. Again, we are unaware of any research specifying or testing these claims.

Empirical Approach

We ran two preregistered[23] online survey experiments designed to test our six hypotheses. The first experiment examined how respondents evaluated prospective

21. Piercey 2009.
22. Tetlock 2005.
23. See Evidence in Governance and Politics at <www.egap.org>.

national security decisions based on probability assessments presented in a series of vignettes. The second asked respondents to make their own probability assessments in response to questions regarding foreign policy issues.

We administered these surveys to a total of 407 national security officials enrolled in two advanced military education programs.[24] Sixty-two percent of these national security officials were active-duty US military officers at the ranks of lieutenant colonel and colonel (for the US Army, Air Force, and Marines), or the equivalent US Navy ranks of commander and captain.[25] Since the US military requires attendance at similar programs for promotion to the rank of colonel or captain, our survey respondents represent a cross-section of military officials.[26] These programs also contained substantial numbers of non-US military officers and civilian officials. Thirteen percent of these respondents were foreign military officers and 25 percent were civilians from the US Intelligence Community, Department of State, and other national-security-related agencies. We describe respondent demographics for individual experiments and in supporting material.

We paired these experiments with surveys of 3,017 respondents via Amazon Mechanical Turk (AMT).[27] Pairing elite and non-elite samples increases statistical power while allowing us to compare how different populations of respondents react to experimental treatments. Generally speaking, elite and non-elite respondents responded to our experimental treatments in similar ways (though the elite sample was substantially more effective at probability assessment on the whole). These complementary results contribute to a growing body of scholarship that indicates how decision science research conducted on non-elite samples can generate plausible insights into national security decision making.[28]

How Decision Makers Interpret Probability Assessments

We tested Hypotheses 1a–1c by presenting respondents with fictional vignettes involving national security decisions.[29] Presented in random order, these scenarios

24. Our larger elite sample was made up of students at the National War College, and our smaller elite sample involved students at Air University. We thank the 407 national security officials in these programs for the time they volunteered to participate in our research.

25. Military institutions distributed our surveys to blocks of students. Response rates exceeded 95 percent.

26. Particularly our larger sample of national security officials drawn from the National War College, which draws students from each of the country's military services.

27. We conducted surveys with our National Security Officials and AMT samples between 5 and 7 August 2015. AMT respondents were required to be US residents at least eighteen years of age. We compensated AMT respondents $2.04 for completing a survey that took an average of seventeen minutes to complete, corresponding to an hourly wage of $7.20, which roughly matches the federal minimum standard. We thank CREATE for its support in providing fair compensation to AMT respondents in this study. On the use of AMT surveys in political science and in international relations specifically, see Berinsky, Huber, and Lenz 2012; Huff and Tingley 2015; Hyde 2015; Renshon 2015; and Williamson 2016.

28. See Dhami et al. 2015 for more discussion of this point.

29. Prominent examples of how international relations scholars have used fictional vignettes to evaluate support for national security decisions include Kertzer and Brutger 2016; Press, Sagan, and Valentino 2013; and Tomz and Weeks 2013.

included a hostage rescue mission, a drone strike, and aiding local security forces in counterinsurgency. The appendix provides examples of the hostage rescue scenario, and supplementary material contains full text of the others. Following each vignette, we asked respondents how strongly they supported the proposed action and how strongly they supported waiting for additional information before deciding. We elicited these evaluations on seven-point scales.[30]

We randomly assigned respondents to qualitative and quantitative assessment conditions. In the qualitative assessment condition, all probability assessments were expressed using one of the seven qualitative terms shown at the top of Figure 1. In the quantitative assessment condition, we converted those qualitative phrases into numeric percentages.[31] We administered this survey to 208 participants in an advanced military education program, which we call our "National Security Officials Sample."[32] We also administered this survey to 1,458 respondents on Amazon Mechanical Turk.[33]

We randomly varied probability assessments within each vignette to represent what we considered to be "optimistic," "neutral," or "pessimistic" information about proposed actions. For example, in the "optimistic" version of the hostage rescue vignette, intelligence analysts estimated that there was an 80 percent chance (or that it was "very likely") that the hostages were at the suspected location. This assessment was placed at 65 percent (or "likely") and 50 percent (or "even chance") for the neutral and pessimistic versions of this scenario, respectively.[34] Providing a check on internal validity, Figure 2 shows how respondents generally opposed proposed actions described with what we considered to be pessimistic assessments, and they generally supported actions described with what we considered to be optimistic assessments. These comparisons prove that respondents consistently used the probabilistic information provided in these vignettes to form their views about proposed actions.

To extend our elite sampling, we also administered a shorter survey, containing only the "neutral" hostage vignette, to 199 students in a second advanced military education program. We refer to this supplementary survey experiment as "Elite Sample B."[35] Altogether, we administered Experiment 1 to 1,857 respondents, who evaluated a total of 5,173 scenarios.

30. We also asked respondents to "write a few sentences" justifying their views and how "confident" they were in making their choices. We discuss the value of examining confidence assessments in note 39.

31. We translated qualitative probability assessments into the numeric percentage closest to the middle of the range that each qualitative phrase represented, rounded to multiples of 0.05. Thus, we converted "even chance" to "50 percent," "likely" to "65 percent," and so forth.

32. This sample was 85 percent male, 82 percent white, and all respondents possessed a college degree. Eighty-seven percent of respondents were US citizens. Seventy-five percent were military officers.

33. This sample was 48 percent male and 80 percent white, with 61 percent of respondents possessing a college degree. Average age was thirty-five.

34. Supporting material shows how probability assessments varied across vignette versions.

35. Eighty-six percent of this sample were US citizens; 78 percent were active-duty military. We were asked not to record information on gender and race in this sample.

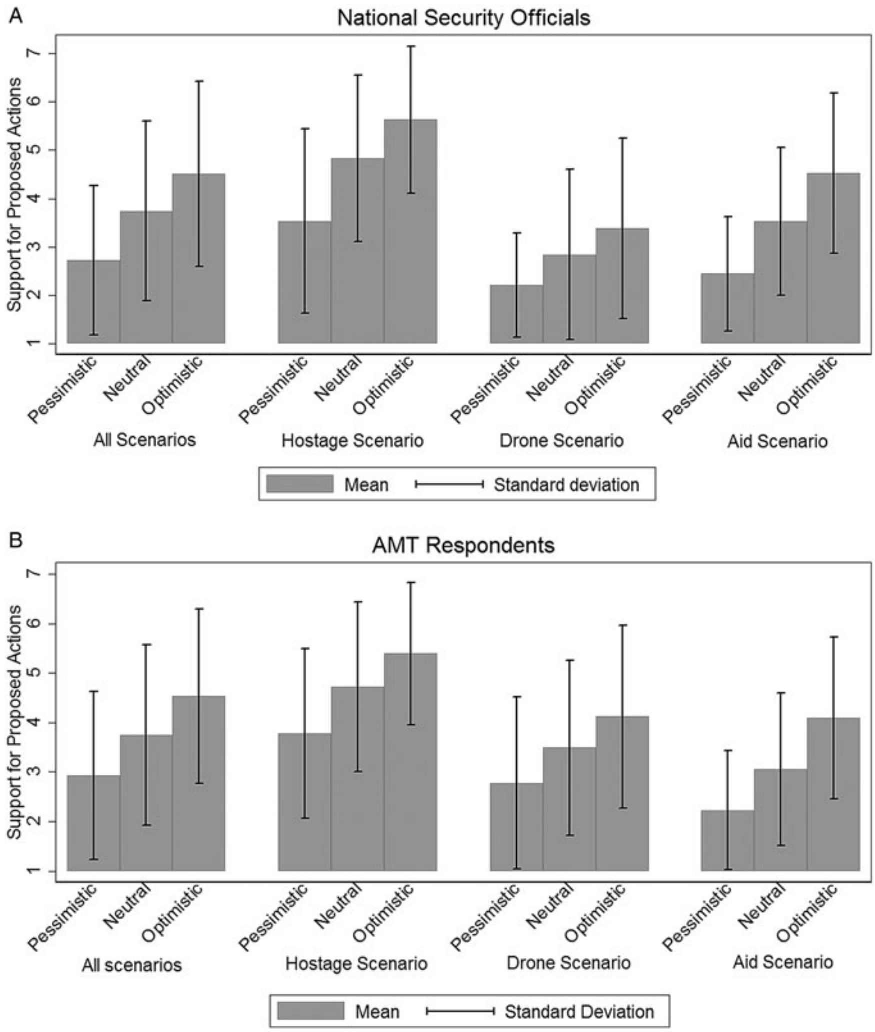


FIGURE 2. *Support for proposed actions across scenarios*

Results

Hypothesis 1a predicts that quantifying probability assessments should make decision makers more likely to support proposed actions. Table 1 examines this hypothesis by comparing respondents' support for proposed actions (measured on a seven-point scale) depending on whether we communicated probability assessments in qualitative versus quantitative form, for each type of vignette that we presented to respondents. In supplementary material, we replicate these patterns using multivariate analyses that control for individual attributes.

TABLE 1. *Support for proposed actions across scenarios*

National Security Officials

		Scenario version			
		Pessimistic	*Neutral*	*Optimistic*	***All***
Scenario topic	*Hostage*	−0.44 ($p = 0.35$)	−0.92 ($p = 0.03$)*	−0.19 ($p = 0.62$)	**−0.32 ($p = 0.24$)**
	Drone	0.13 ($p = 0.63$)	−0.55 ($p = 0.23$)	0.16 ($p = 0.73$)	**−0.08 ($p = 0.74$)**
	Aid	0.23 ($p = 0.42$)	0.19 ($p = 0.61$)	−0.36 ($p = 0.38$)	**0.09 ($p = 0.72$)**
	All	**−0.10 ($p = 0.63$)**	**−0.39 ($p = 0.13$)**	**−0.04 ($p = 0.88$)**	**−0.10 ($p = 0.49$)**

AMT Respondents

		Scenario version			
		Pessimistic	*Neutral*	*Optimistic*	*All*
Scenario topic	*Hostage*	−0.78 ($p < 0.001$)***	−0.98 ($p < 0.001$)***	−0.29 ($p = 0.02$)*	−0.74 ($p < 0.001$)***
	Drone	−0.38 ($p = 0.02$)*	−0.86 ($p < 0.001$)***	0.18 ($p = 0.27$)	−0.35 ($p < 0.001$)***
	Aid	−0.13 ($p = 0.22$)	−0.45 ($p = 0.001$)**	−0.27 ($p = 0.07$)	−0.26 ($p = 0.002$)**
	All	**−0.39 ($p < 0.001$)*** **	**−0.75 ($p < 0.001$)*** **	**−0.16 ($p = 0.09$)**	**−0.45 ($p < 0.001$)*** **

Notes: This table presents the average change in support for proposed actions, as measured on a seven-point scale, associated with quantifying probability estimates. Estimates reflect ordinary least squares regressions. For the row of results pertaining to "all scenario topics," where there are multiple data points per respondent, regressions include respondent fixed effects and standard errors clustered by respondent. *$p < .05$; **$p < .01$; ***$p < .001$.

Table 1 shows that quantifying probability estimates did not make respondents more likely to support proposed actions. Across all three versions of all three scenarios, we found no instance where quantifying probability assessments consistently increased support for proposed actions. In fact, we found that respondents were on the whole *less* likely to support actions described with numeric probabilities. This difference is statistically significant at the $p < 0.001$ level both in the AMT data and in Elite Sample B.[36]

Hypothesis 1b predicts that quantifying probability assessments amplifies policy evaluations. If this is true, we should observe two findings. First, quantifying probability assessments should depress respondents' support for proposed actions most extensively in pessimistic scenarios, indicating that analytic precision makes bad options seem worse. Second, quantifying probability assessments should *increase* respondents' support for proposed actions in *optimistic* scenarios, indicating that analytic precision amplifies positive cues. Yet Table 1 shows how our experiment refuted both of these predictions. While we found that quantifying probability assessments did reduce respondents' support for proposed actions in pessimistic scenarios, we found that this treatment effect was actually greatest in *neutral* scenarios.

36. See supplementary material for full analysis of Elite Sample B. Mean (standard deviation) support for acting in the qualitative assessment condition was 5.33 (1.56), compared to 4.52 (1.86) in the quantitative assessment condition.

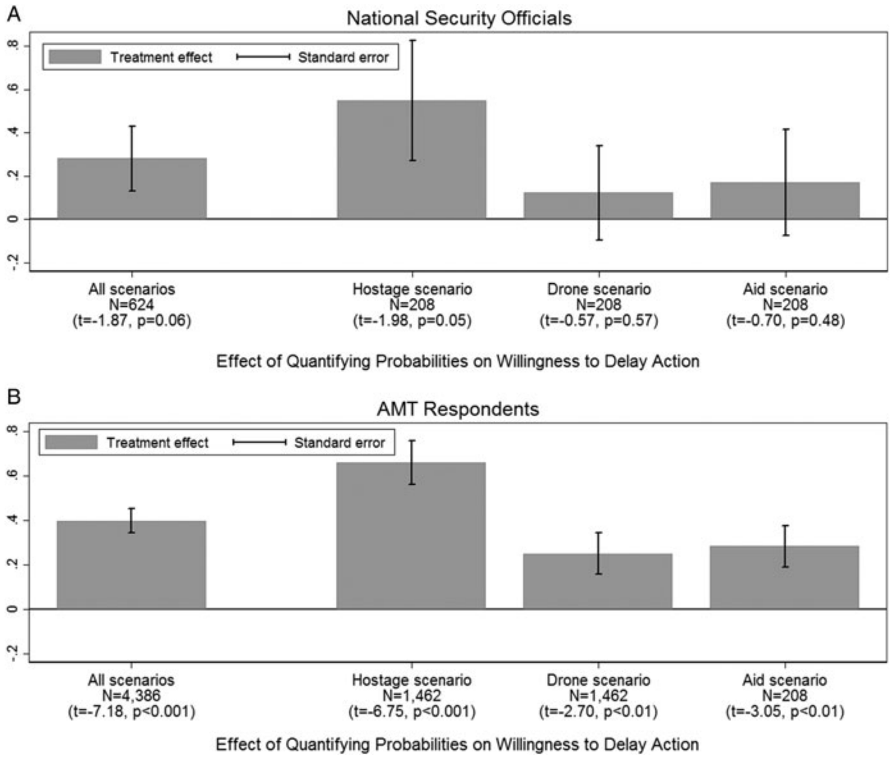


FIGURE 3. *How quantifying probabilities influenced willingness to delay action*

Moreover, we found no evidence that quantifying probability assessments increased support for proposed actions in optimistic scenarios.[37]

Hypothesis 1c predicts that quantifying probability assessments should make decision makers less willing to delay action to gather additional information. Our results refute this hypothesis, as well. Figure 3 shows that respondents presented with quantitative probability assessments were, in fact, *more* willing to gather additional information. These results were even stronger among the 199 respondents in Elite Sample B, for whom mean support for delaying decision across respondents was nearly a full point higher in the quantitative assessment condition.[38]

Altogether, the results from our first survey experiment do not suggest that quantitative and qualitative probability assessments are interchangeable. Respondents given numeric probabilities were more cautious in supporting proposed actions.

37. Supplementary material again replicates this finding via multivariate analysis.

38. Support for delaying action in Elite Sample B was 3.14 for the qualitative assessment condition (standard deviation 1.97), and 4.11 in the quantitative assessment condition (standard deviation 2.07), $p = 0.001$.

This is not necessarily ideal: sometimes, the right move is to act, incomplete information and all.[39]

Nevertheless, a large body of scholarship supports a general consensus that national security decision makers often neglect to address key uncertainties, and that they can be overly inclined towards taking risks.[40] One argument in favor of making probability assessments explicit is that they can prevent decision makers from glossing over key uncertainties or interpreting ambiguous information in ways that support excessively risky behavior. The illusions-of-rigor argument is important because it suggests that attempts to highlight uncertainty by making probability assessments more precise can backfire, unintentionally increasing decision makers' willingness to take risks on the basis of incomplete information. Yet the data presented here support the opposite argument, roundly disconfirming three plausible concerns about the drawbacks of probabilistic precision.

How Analysts Estimate Probabilities

We tested Hypotheses 2a–2c by asking respondents to make probability estimates in response to thirty-five randomly ordered questions about foreign policy and national security. Thirty questions had factual, yes-or-no answers. (For example, "In your opinion, what are the chances that Russia's economy grew in 2014?"). Five questions involved forecasts (for example, "In your opinion, what are the chances that within the next six months, Syrian President Bashar al-Assad will be killed or no longer living in Syria?").[41] We randomly assigned respondents to estimate probabilities using either numeric percentages or the seven qualitative expressions shown at the top of Figure 1. We administered this survey to our National Security Officials sample and to 1,561 respondents on Amazon Mechanical Turk.[42] These surveys produced 61,901 probability estimates.[43]

39. One interpretation for these results is that respondents simply found numeric probabilities more confusing. We anticipated this possibility when constructing our survey, and thus asked each respondent to rate their level of "confidence" in making their decisions on a seven-point scale. We found that respondents were in fact slightly *more* confident in their ability to evaluate uncertainty when presented with numeric probabilities, though this finding was not statistically significant.

40. See Johnson 2004; Kahneman and Renshon 2007; and Rapport 2015.

41. We also asked five questions involving statements about current or previous states of the world that were unverifiable at the time of the survey. (For example, "In your opinion, what are the chances that high-ranking members of Pakistan's intelligence services knew that Osama bin Laden was hiding in Abbottabad?") Supporting information shows that analysts provided similar distributions of responses for these questions as on other items in the survey.

42. Respondents in our National Security Officials sample took surveys containing both Experiments 1 and 2. We assigned these respondents to the same treatment condition across experiments and randomized the order in which these experiments appeared. AMT respondents were randomly assigned to complete only one of our two experiments, hence the uneven sample sizes. Our AMT respondents were 48 percent male and 81 percent white. Their average age was thirty-five. Sixty-one percent had a college degree.

43. We dropped fourteen estimates because they were greater than 100 percent, presumably a result of typographical errors.

We scored qualitative and quantitative estimates using the following procedure (though we note how our results are robust to several scoring methods). First, we calculated the mean numeric assessment corresponding to each word of estimative probability for each question we posed. We then replaced every qualitative assessment in the data set with those question-word-specific means. We replaced every quantitative assessment with those means as well. Otherwise, quantitative assessments could have exhibited greater variance, which would prevent scoring qualitative and quantitative estimates on a level playing field.

After translating probability estimates into equivalent terms, we evaluated their accuracy using Brier Scores.[44] Using this method, we found that 81 percent of AMT respondents and 98 percent of national security officials provided assessments that were more informative, on average, than random guessing.[45] This indicates that a large majority of participants took the probability estimation exercise seriously, especially given how subject-matter experts often struggle to beat the "as-good-as-random" standard when evaluated with proper scoring rules.[46]

Results

Figure 4 compares cumulative distributions of respondents' mean Brier scores. When respondents estimated probabilities numerically, their responses were less accurate on average than when using "words of estimative probability." Among National Security Officials, the disparity between average Brier scores for quantitative and qualitative assessors was 14 percent. Among AMT respondents, the equivalent gap was 11 percent. Both of these comparisons were significant at the $p < .001$ level.

Alternative scoring rules produced similar patterns. With logarithmic scoring, the difference across treatment conditions in respondents' mean Brier Scores was 17 percent for National Security Officials and 15 percent for AMT respondents.[47] If we round probability estimates to the midpoint of each "word of estimative probability," instead of using question-specific interpretations as described earlier, then the gap in performance using Brier Scores was 11 percent for National Security Officials and 10 percent for AMT respondents.[48]

Table 2 shows how respondents using numeric probabilities substantially underperformed compared to respondents using words of estimative probability. This

44. Brier Scores compute the mean squared error of a probabilistic assessment. Thus, if a respondent assigns probability p to a statement that proves true, then the outcome is assigned a value of 1 and the respondent's Brier Score for that prediction is $(1-p)^2$. If the statement proves false, then the respondent's Brier Score for that prediction is $(0-p)^2$.

45. Randomly assigning probabilities, with a uniform distribution, to questions with binary outcomes, generates an expected Brier Score of 0.335.

46. Tetlock 2005.

47. Logarithmic scoring pays the natural logarithm of the probability respondents assigned to the "correct" answer. We replaced estimates of 0.00 and 1.00 with 0.01 and 0.99, respectively, otherwise logarithmic payoffs can return infinitely negative scores.

48. See supporting information for additional analysis.

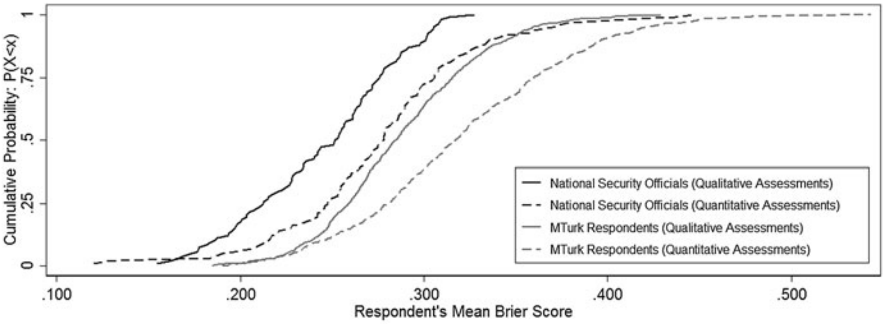


FIGURE 4. *Cumulative distributions of respondent Brier scores by treatment group*

table presents ordinary least squares regressions predicting the accuracy of each judgment in our data set. Each model included fixed effects for respondents and questions. We also clustered standard errors by respondent. In addition to capturing whether or not a respondent assessed probabilities using numbers, we controlled for several individual attributes: a four-point measure of respondent numeracy,[49] along with indicators for whether respondents were female, US citizens, military officers, and native

TABLE 2. *Relationship between quantitative assessment, respondent certitude, and judgmental accuracy*

	National Security Officials		*AMT Respondents*	
	Model 1	*Model 2*	*Model 3*	*Model 4*
QUANTITATIVE ASSESSMENT	0.032 (.01)***	0.001 (.01)	0.013 ($2.1e^{-3}$)***	−0.002 ($2.3e^{-3}$)
CERTITUDE	0.032 (.02)	−0.040 (.03)	0.297 (.01)***	0.256 (.01)***
QUANTITATIVE ASSESSMENT × CERTITUDE		0.124 (.04)***		0.070 (.02)***
FORECAST	0.148 (.02)***	0.146 (.02)***	0.241 (.01)***	0.241 (.01)***
NUMERACY	−0.003 ($2.5e^{-3}$)	−0.003 ($2.5e^{-3}$)	−0.004 ($9.7e^{-4}$)***	−0.004 ($9.6e^{-4}$)***
MILITARY SERVICE	−0.001 (.01)	−0.001 (.01)	0.015 (.02)	0.013 (.02)
FEMALE	0.003 (.01)	0.003 (.01)	0.009 ($2.2e^{-3}$)***	0.009 ($2.2e^{-3}$)***
US CITIZEN	0.011 (.01)	0.015 (.01)	0.003 (.01)	0.002 (.01)
ENGLISH NATIVE LANG.	−0.020 (.01)	−0.022 (.01)	0.003 (.01)	0.003 (.01)
EDUCATION LEVEL			−0.011 ($1.6e^{-3}$)***	−0.011 ($1.6e^{-3}$)***
Constant	0.097 (.02)***	0.113 (.02)***	0.218 (.01)***	0.227 (.01)***
N	7,280	7,280	54,621	54,621
R^2	0.218	0.219	0.221	0.222

Notes: Table 2 presents ordinary least squares regressions predicting Brier scores for individual probability assessments. Standard errors clustered by respondent. Fixed effects for respondents and questions not shown. Note that higher Brier Scores indicate less accurate assessments. *p < .05; **p < .01; ***p < .001.

49. Cokely et al. 2012.

English speakers.[50] For AMT respondents we also included a three-point measure of education.[51] Models 1 and 3 demonstrate that, even when controlling for these attributes, respondents assessing numeric probabilities obtained consistently higher (that is, worse) Brier Scores.

One limitation of our approach is that it assumes respondents actually followed the "words of estimative probability" lexicon they were asked to use. For example, given that this lexicon divides the number line into seven equal bins, the lowest term (*remote*) covers estimates of 0 to 14 percent. But respondents may be inclined to use the term *remote* only for very low probabilities. If so, then respondents might have *intended* for the term *very likely* to cover probabilities that were substantially smaller than the way we interpreted these estimates.

To examine whether our results hinge on this issue, we replicated our analysis in two ways.[52] First, we scored qualitative estimates according to Mosteller and Youtz's metastudy of how respondents typically evaluate these terms,[53] and we rounded numeric estimates to the nearest of these anchors.[54] This method found that quantitative assessors produced mean Brier scores that were 6 percent worse among National Security Officials ($p < 0.01$) and 4 percent worse among AMT respondents ($p < 0.001$).

Next, we replicated our original analysis as if respondents had used the alternative words of estimative probability spectrum defined by the US Director of National Intelligence (see Figure 1), under which the terms *remote* and *almost certain* span smaller ranges. This approach showed that the degradations in performance associated with quantitative estimation were 8 and 5 percent among National Security Official and AMT respondents, respectively, with both differences retaining statistical significance at the $p < 0.001$ level.[55] Thus while it is important to acknowledge the difficulty of evaluating qualitative probability assessments even when instructing respondents to make those estimates according to structured lexicons—and this is one clear drawback that all such lexicons share—this issue does not appear to have driven our results.[56]

50. In the MTurk sample, the MILITARY SERVICE variable indicates respondents with any current or previous military service. Fewer than 1 percent of MTurk respondents were active-duty military personnel.

51. We measured education on a three-point scale: high-school diploma or less, college degree (two- or four-year), and postgraduate training. All elite respondents had a college degree and some postgraduate training.

52. See supporting information for additional analysis.

53. Mosteller and Youtz 1990.

54. Thus we translated the "almost certain" to 86 percent, "very likely" to 85 percent, "likely" to 69 percent," "even chance" to 50 percent," "unlikely" to 16 percent, "very unlikely" to 8 percent," and "remote" to 3 percent. Mosteller and Youtz do not examine the word *remote*, so we used the 3 percent figure they assign to "almost never."

55. See supplementary material for more information.

56. An intriguing possibility is that respondents were better calibrated on qualitative assessments because they were *not* using these assessments as intended: that the lower proportion of assessments assigned to "remote" or "almost certain" probabilities represents a misuse of those terms as opposed to more accurate judgments.

Precision and Certitude

Figure 5 compares the frequency with which qualitative and quantitative assessors provided estimates corresponding to each word of estimative probability. Respondents who used numeric probabilities were substantially more willing to make estimates at the extreme ends of the probability scale.[57] These data suggest support for Hypothesis 2c, that quantifying probability estimates makes analysts more likely to offer assessments with excessive certitude.

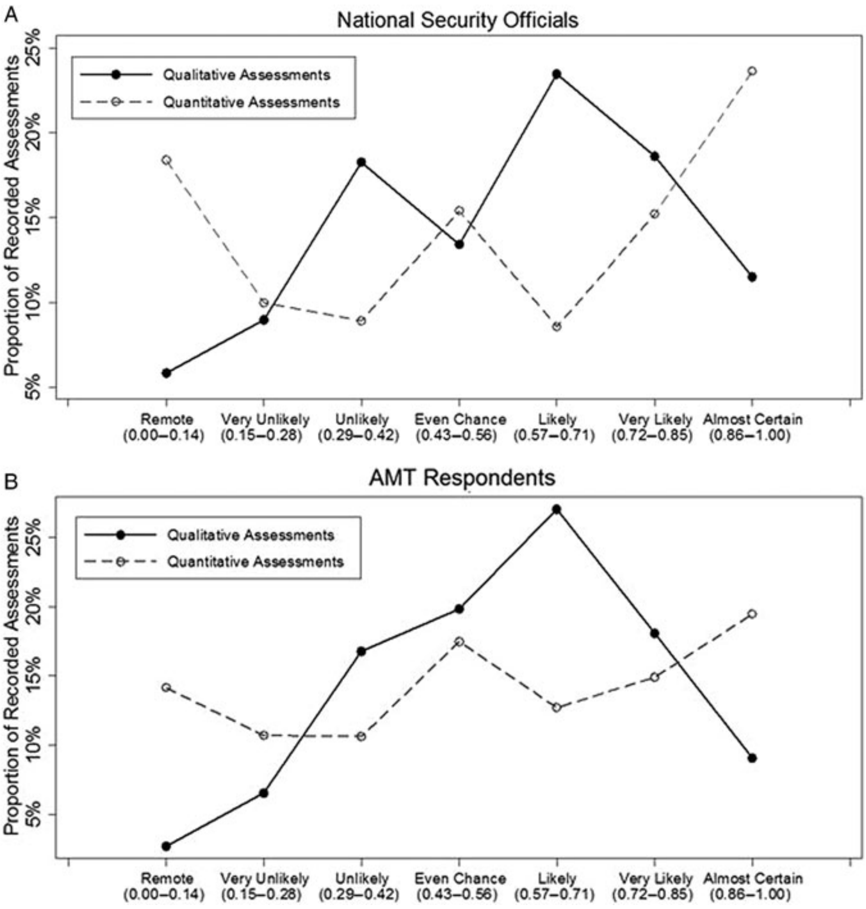


FIGURE 5. *Comparing distributions of qualitative/quantitative probability assessments*

57. All differences in proportions between qualitative and quantitative assessors in Figure 5 are statistically significant at $p < 0.01$. See supporting material for further analysis.

To evaluate this hypothesis, we defined each probability estimate's CERTITUDE as the absolute value of its difference from 50 percent.[58] Then we reran our previous analysis examining the predictors of judgmental accuracy. In these new models, however (see Table 2, Models 2 and 4) we added an interaction term between the dummy variable indicating that a respondent used numeric probabilities, and a measure of how certain a respondent's judgment was.

These models show a statistically significant ($p < 0.001$) interaction term between CERTITUDE and QUANTITATIVE ASSESSMENT, confirming that when numeric assessors made estimates with greater certainty, this systematically degraded their accuracy. Moreover, including this interaction term eliminates the disparity we previously observed between QUANTITATIVE ASSESSMENT and respondents' Brier Scores. By contrast, we found no indication that the performance gap between qualitative and quantitative assessors was driven by respondents' numeracy, gender, language, nationality, age, education, or military experience. Interaction terms between QUANTITATIVE ASSESSMENT and these factors were statistically insignificant when added to the models presented in Table 2, and doing so did not influence other results.

Of the subsets of respondents who participated in our study, the one for whom the decrement associated with quantitative assessment is by far the largest is the worst-performing assessors. For example, if we exclude from the analysis respondents whose average Brier Scores fell into the bottom quartile of their respective samples, then the accuracy reduction associated with quantitative assessment declines to 5 percent among National Security Officials ($p < 0.05$) and to 3 percent among AMT respondents ($p < 0.001$). If we limit the analysis to respondents whose Brier Scores were better than the median within their respective samples, then there is no statistically significant difference between the performance of qualitative and quantitative assessors.[59] This finding raises the possibility that the numbers-as-a-second-language problem mainly appears among respondents with low levels of motivation, and that proper training or more effort would substantially mitigate this bias.

Directions for Future Research

This study addresses long-standing debates about the desirability of probabilistic precision in high-stakes decision making, particularly in the national security domain. Our main theoretical contribution was to distill behavioral concerns about this

58. National Security Officials' estimates had an average certitude of 0.26 in the quantitative assessment condition versus 0.19 in the qualitative assessment condition. For AMT respondents, average certitude was 0.29 and 0.22 in these conditions, respectively. Both differences are statistically significant at the $p < 0.001$ level.

59. AMT assessors in the quantitative assessment condition return slightly better Brier Scores ($p = 0.41$), while the degradation in performance among numerical assessors in the National Security Officials sample is not statistically significant ($p = 0.31$).

practice into six falsifiable hypotheses. We evaluated those hypotheses with two preregistered survey experiments administered to paired samples of national security professionals and respondents from Amazon Mechanical Turk. To our knowledge, this is the first attempt to submit long-standing skepticism about quantifying probability assessments in national security to direct empirical testing, let alone to do so with an elite sample.

Our results do not provide a clear-cut victory either for proponents or for skeptics of quantifying probability assessments. Our first experiment roundly rejected claims that numeric probabilities create illusions of rigor that goad decision makers into supporting proposed actions on the basis of incomplete information. However, our second experiment indicated that quantifying probabilities led respondents to provide judgments with excessive certitude, particularly among low-quality assessors. Thus the study's main contribution is to advance basic research and not a policy prescription. Specifically, our empirical findings hone a wide-ranging list of concerns into a specific and previously undocumented bias.

An important question for subsequent research is whether it is feasible to correct the bias we identified, and previous scholarship suggests this is the case. For example, the Good Judgment Project has shown that even one-hour training sessions in probability assessment can markedly reduce foreign policy analysts' tendencies to assess uncertainty with excessive certitude.[60] Similarly, when Mandel and Barnes calibrated a large volume of intelligence estimates, they found analysts' judgments to be *under*confident on the whole.[61] Especially because the treatment effect we observed in our study originated mainly with our worst assessors, we suspect that our results overstate the extent to which professional analysts, who possess greater training, experience, and incentives for careful reasoning, would suffer from this problem. On balance, we therefore believe that our results support proponents of quantifying probability assessments in national security, so long as those proponents do not claim that probabilistic precision is a free lunch.[62] Nevertheless, this is a proposition that demands rigorous analysis of its own.

Our results suggest four further directions for additional research. First, scholars should explore why quantifying probability assessments reduces support for proposed actions. In particular, it is important to understand whether this additional caution represents more careful considerations of risk as opposed to a bias against basing decisions on numerical judgments. Second, our empirical results call for further study of why respondents attach greater certainty to quantitative probability assessments, especially whether this finding is an artifact of elicitation scales as opposed to reflecting different cognitive processes. Third, scholars can extend

60. Mellers et al. 2014.

61. Mandel and Barnes 2014.

62. Of course, systems for expressing qualitative probabilities are no free lunch either. The system used by the Defense Intelligence Agency alone (see Figure 1) requires analysts and decision makers to memorize the meanings of thirty-four probabilistic terms, and to understand how different agencies use similar terms in different ways.

similar research into other disciplines, such as medicine, law, regulation, and climate science, that feature their own prominent debates about the costs and benefits of probabilistic precision. While our experimental findings do not directly apply to fields outside of national security decision making, our basic method for specifying and testing key concerns about probabilistic precision can be used in nearly any area of high-stakes decision making.

Fourth, and most broadly, we seek to reorient debates about probabilistic precision from epistemology to empirics. Most published objections to analytic precision in national security or in other fields revolve around claims about what kinds of language seem most appropriate for conveying the inherently subjective nature of world politics. For many scholars and practitioners, expressing subjective judgments precisely simply feels wrong. But ultimately, if probabilistic precision threatens analysis and decision making, then that should have observable empirical consequences. If not, then intuitive discomfort with quantitative expression is a poor basis for leaving key assumptions deliberately vague. This study offers evidence, consistent across both elite and non-elite samples, sharpening claims about which behavioral consequences of probabilistic precision appear to be most problematic, under what conditions they are most likely to occur, and how scholars can place this debate on a sounder empirical footing.

Supplementary Material

Supplementary material for this research note is available at <https://doi.org/10.1017/S0020818317000352>.

Appendix

This appendix displays how the hostage rescue scenario appeared to respondents in both the qualitative and quantitative treatment conditions. We present material from the "neutral" version of that scenario. Supplementary material contains further descriptions of the other vignettes we employed in our survey experiments.

Hostage Scenario, Neutral Version, Qualitative Assessments

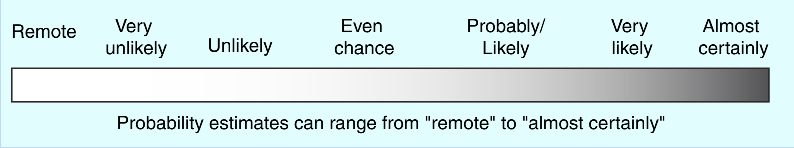
ISIS is holding three American aid workers hostage. The US Intelligence Community has used human intelligence and communications intercepts to trace these hostages to a compound in Manbij, Syria.

Intelligence analysts stress that their judgments are subjective and that they are based on incomplete information. However, after reviewing all available information, they estimate that it is likely that the hostages are at the Manbij compound. US Special Forces have designed and rehearsed a hostage rescue mission. Based on their track record and on the specific details of this plan, military officials assess that if the hostages are in this location, it is very likely that Special Forces can retrieve the hostages alive.

This mission entails several risks. Analysts believe there is an even chance that ISIS will wound or kill US soldiers on this mission. They believe that it is possible, though unlikely, that the mission would inadvertently wound or kill a small number of innocent civilians living near the suspected compound. They also warn that if the raid fails (including if the aid workers are not being held in the Manbij location), ISIS will almost certainly execute the hostages.

Summary of estimated chances:

- The hostages are at the Manbij compound: *Likely*
- If the hostages are in this location, Special Forces can retrieve them alive: *Very likely*
- ISIS will wound or kill US soldiers on this mission: *Even chance*
- The mission would inadvertently wound or kill innocent civilians: *Unlikely*
- ISIS will kill the hostages if the raid fails: *Almost certainly*



Hostage Scenario, Neutral Version, Quantitative Assessments

ISIS is holding three American aid workers hostage. The US Intelligence Community has used human intelligence and communications intercepts to trace these hostages to a compound in Manbij, Syria.

Intelligence analysts stress that their judgments are subjective and that they are based on incomplete information. However, after reviewing all available information, they estimate that there is a 65 percent chance that the hostages are at the Manbij compound. US Special Forces have designed and rehearsed a hostage rescue mission. Based on their track record and on the specific details of this plan, military officials assess that if the hostages are in this location, there is an 80 percent chance that Special Forces can retrieve the hostages alive.

This mission entails several risks. Analysts believe there is 50 percent chance that ISIS will wound or kill US soldiers on this mission. They believe that there is a 35 percent chance that the mission would inadvertently wound or kill a small number of innocent civilians living near the suspected compound. They also warn that if the raid fails (including if the aid workers are not being held in the Manbij location), there is a 95 percent chance that ISIS will execute hostages.

Summary of estimated chances:

- The hostages are at the Manbij compound: *65 percent*
- If the hostages are in this location, Special Forces can retrieve them alive: *80 percent*
- ISIS will wound or kill US soldiers on this mission: *50 percent*
- The mission would inadvertently wound or kill innocent civilians: *35 percent*
- ISIS will kill the hostages if the raid fails: *95 percent*

References

Barnes, Alan. 2015. Making Intelligence Analysis More Intelligent. *Intelligence and National Security* 31 (1):327–44.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20 (3):351–68.

Betts, Richard K. 2006. *Enemies of Intelligence: Knowledge and Power in American National Security*. New York: Columbia University Press.

Beyerchen, Alan. 1992/93. Clausewitz, Nonlinearity, and the Unpredictability of War. *International Security* 17 (3):59–90.

Budescu, David V., Han-Hui Por, Stephen B. Broomell, and Michael Smithson. 2014. The Interpretation of IPCC Probabilistic Statements Around the World. *Nature Climate Change* 4: 508–12.

Cokely, Edward T., Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-Retamero. 2012. Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making* 7 (1):25–47.

Dhami, Mandeep K. 2013. *Understanding and Communicating Uncertainty in Intelligence Analysis*. Report Prepared for Her Majesty's Government. London, UK.

Dhami, Mandeep K., David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. Improving Intelligence Analysis with Decision Science. *Perspectives on Psychological Science* 10 (6):753–57.

Ellsberg, Daniel. 1961. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics* 75 (4): 643–69.

Fingar, Thomas. 2011. *Reducing Uncertainty: Intelligence and National Security*. Stanford, CA: Stanford Security Studies.

Friedman, Jeffrey A., and Richard Zeckhauser. 2015. Handling and Mishandling Estimative Probability. *Intelligence and National Security* 30 (1):77–99.

Gigerenzer, Gerd. 2002. *Calculated Risks: How to Know When Numbers Deceive You*. New York: Simon and Schuster.

Hafner-Burton, Emilie M., Stephan Haggard, David A. Lake, and David G. Victor. 2017. The Behavioral Revolution and the Study of International Relations. *International Organization* 71 (S1):S1–S31.

Heuer, Richards J., Jr. 1999. *Psychology of Intelligence Analysis*. Washington, DC: Center for the Study of Intelligence.

Huff, Connor, and Dustin Tingley. 2015. "Who Are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents. *Research and Politics* 2 (3): 1–12.

Hyde, Susan D. 2015. Experiments in International Relations: Lab, Survey, and Field. *Annual Review of Political Science* 18 (1):403–24.

Jervis, Robert. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, NJ: Princeton University Press.

———. 2010. *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War*. Ithaca, NY: Cornell University Press.

Johnson, Dominic D.P. 2004. *Overconfidence and War: The Havoc and Glory of Positive Illusions*. Cambridge, MA: Harvard University Press.

Johnston, Rob. 2005. *Analytic Culture in the US Intelligence Community*. Washington, DC: Center for the Study of Intelligence.

Kahneman, Daniel, and Jonathan Renshon. 2007. Why Hawks Win. *Foreign Policy* 158:34–38.

Kent, Sherman. 1964. Words of Estimative Probability. *Studies in Intelligence* 8 (4):49–65.

Kertzer, Joshua D., and Ryan Brutger. 2016. Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory. *American Journal of Political Science* 60 (1):234–49.

Lanir, Zvi, and Daniel Kahneman. 2006. An Experiment in Decision Analysis in Israel in 1975. *Studies in Intelligence* 50 (4). Available at <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol50no4/an-experiment-in-decision-analysis-in-israel-in-1975.html>.

Levy, Jack S. 2013. Psychology and Foreign Policy Decision-Making. In *The Oxford Handbook of Political Psychology*, 2nd ed., edited by Leonie Huddy, David O. Sears, and Jack S. Levy, 301–33. Oxford, UK: Oxford University Press.

Lowenthal, Mark M. 2006. *Intelligence: From Secrets to Policy*, 3rd ed. Washington, DC: CQ Press.

Mandel, David R., and Alan Barnes. 2014. Accuracy of Forecasts in Strategic Intelligence. *Proceedings of the National Academy of Sciences* 111 (30):10984–89.

McChrystal, Stanley. 2009. *COMISAF Initial Assessment*. Kabul, Afghanistan: Headquarters, International Security Assistance Force.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, Don Moore, Pavel Atanasov, Samuel A. Swift, Terry Murray, Eric Stone, and Philip E. Tetlock.

2014. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science* 25 (5):1106–15.

Mintz, Alex, and Nehemia Geva, eds. 1997. *Decisionmaking on War and Peace: The Cognitive-Rational Debate*. Boulder, CO: Lynne Rienner.

Mosteller, Frederick, and Cleo Youtz. 1990. Quantifying Probabilistic Expressions. *Statistical Science* 5 (1):2–12.

Piercey, M. David. 2009. Motivated Reasoning and Verbal vs. Numerical Probability Assessment: Evidence from an Accounting Context. *Organizational Behavior and Human Decision Processes* 108 (2):330–41.

Press, Daryl G., Scott D. Sagan, and Benjamin A. Valentino. 2013. Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-Use of Nuclear Weapons. *American Political Science Review* 107 (1):188–206.

Rapport, Aaron. 2015. *Waging War, Planning Peace: US Noncombat Operations and Major Wars*. Ithaca, NY: Cornell University Press.

Renshon, Jonathan. 2015. Losing Face and Sinking Costs: Experimental Evidence on the Judgment of Political and Military Leaders. *International Organization* 69 (3):659–95.

Rovner, Joshua. 2011. *Fixing the Facts: National Security and the Politics of Intelligence*. Ithaca, NY: Cornell University Press.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: Wiley.

Shapiro, Jacob N., and Dara Kay Cohen. 2007. Color Blind: Lessons from the Failed Homeland Security Advisory System. *International Security* 32 (2):121—54.

Sunstein, Cass R. 2014. *Valuing Life: Humanizing the Regulatory State*. Chicago, IL: University of Chicago Press.

Tetlock, Philip E. 2005. *Expert Political Judgment*. Princeton, NJ: Princeton University Press.

Tetlock, Philip E., and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. New York: Crown.

Tillers, Peter, and Jonathan Gottfried. 2006. "Case Comment—*United States v. Copeland*." *Law, Probability, and Risk* 5 (2):135–57.

Tomz, Michael R., and Jessica P. Weeks. 2013. Public Opinion and the Democratic Peace. *American Political Science Review* 107 (4):849–65.

US Army. 1997. *Field Manual 101-5: Staff Organization and Operations*. Washington, DC: Department of the Army.

———. 2009. *Field Manual 5-0: The Operations Process*. Washington, DC: Department of the Army.

Wallsten, Thomas. 1990. Costs and Benefits of Vague Information. In *Insights in Decision Making*, edited by Robin M. Hogarth, 28–43. Chicago, IL: University of Chicago Press.

Williamson, Vanessa. 2016. On the Ethics of Crowdsourced Research. *PS: Political Science and Politics* 49 (1):77–81.

Zimmer, Alf C. 1984. A Model for the Interpretation of Verbal Predictions. *International Journal of Man-Machine Studies* 20 (1):121–34.