

Behavioral Consequences of Probabilistic Precision

Experimental Evidence from National Security Professionals

Jeffrey A. Friedman, *Assistant Professor of Government, Dartmouth College*
Jennifer S. Lerner, *Professor of Public Policy and Management, Harvard University*
Richard Zeckhauser, *Frank P. Ramsey Professor of Political Economy, Harvard University*

Working Paper: August 2016
8,534 words

All comments welcome: jeffrey.a.friedman@dartmouth.edu

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 which training may be able to correct.

Acknowledgments:

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, 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. Our work was funded by the Department of Homeland Security (DHS) via the University of Southern California's Center for Risk and Economic Analysis of Terrorism Events (CREATE). 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.

Introduction

Though 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 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 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 U.S. intelligence analysts to describe uncertainty using qualitative language.³ U.S. 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”) which lack clear probabilistic equivalents.⁵ Similar debates about communicating uncertainty surround the conduct of national security analysis in Britain, Canada,

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.

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).

⁴ U.S. Army 2009, pars. 2-19, B-173; U.S. Army 1997, par. 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.

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

[Figure 1]

Many observers find this behavior problematic.⁸ After all, probability is an inherently quantitative concept, representing values between zero and one hundred 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.

⁶ Dhami 2013, Barnes 2015, Lanir and Kahneman 2006.

⁷ Tetlock and Gardner 2015.

⁸ Controversy over this subject dates back to Kent 1964.

⁹ Savage 1954.

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-stake decisions.¹⁰ This scholarship is important because it suggests that if analysts and decision makers scrutinized their judgments more carefully, then 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 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.

The remainder of this study hones the illusions of rigor and numbers as a second language arguments into six testable hypotheses, and then tests these hypotheses through preregistered survey experiments administered to 407 national security professionals and to 3,017 respondents from Amazon Mechanical Turk. Our results roundly contradicted the illusions of rigor argument. Contrary to conventional wisdom, we found that quantifying probability assessments made decision makers less willing to support proposed actions and more receptive to gathering additional

¹⁰ For a recent review of this literature, see 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.

information. Yet our results also supported a particular version of the numbers as a second language argument. We found that when respondents estimated probabilities themselves, quantification magnified 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, which 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 just 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. The last section of this study concludes by suggesting how further scholarship can build on these findings not just in national security but also in other areas of high-stakes decision making.

Section 1. 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

¹² See Gigerenzer 2002, Tillers and Gottfried 2005, Sunstein 2014, and Budescu et al. 2014.

¹³ Beyerschen 1992/93, Jervis 1997, Betts 2006, Fingar 2011, Rovner 2011.

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 such judgments would imply more rigor than we intend.” 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

¹⁴ Lowenthal 2006, 129.

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.¹⁵

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.¹⁶ 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 due to 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 makers' willingness to support proposed actions, and this is our first hypothesis for evaluating the illusions of rigor thesis.

Hypothesis 1a: 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

¹⁵ On the Iraq WMD Estimate's use of evidence, see Betts 2006 and Jervis 2010.

¹⁶ Ellsberg 1961. In other words ambiguity averse decision makers have lower probabilistic thresholds for taking gambles based on probabilities that are more precise.

view, probability assessments are just one of many cues that decision makers use in order 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, then this would represent a second way in which illusions of rigor could impair the quality of national security decision making.

Hypothesis 1b: 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.¹⁷ If quantifying probability estimates leads decision makers to believe that these judgments are more reliable than they really are, then this could also cause 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.

Hypothesis 1c: 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. Yet 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.

¹⁷ On information acquisition and foreign policy decision making, see Mintz and Geva 1997, Heuer 1999, Friedman and Zeckhauser 2015.

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,¹⁸ and many national security analysts are said to be especially uncomfortable expressing subjective judgments using numbers.¹⁹ 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.

Hypothesis 2a: 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.

¹⁸ Zimmer 1984, Wallsten 1990.

¹⁹ Kent 1964, Johnston 2005.

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 this allows them to redefine their judgments after the fact in a manner that deflects criticism, a practice known as “elastic redefinition.”²⁰ Because explicit probability assessments foreclose opportunities for elastic redefinition, this could exacerbate analysts’ concerns about receiving criticism when their judgments appear to be mistaken. In order to reduce this prospective criticism, analysts might choose to express their views with less certainty than these judgments deserve.

Hypothesis 2b: Quantifying probability estimates causes analysts to attach less certainty to their judgments.

Yet 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,²¹ 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 ninety or ninety-five percent. Using the guidelines shown 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 eighty

²⁰ Piercey 2009.

²¹ Tetlock 2005.

percent. In this way, coarsening probability assessments could play a helpful role in mitigating analysts' tendencies to make judgments with unjustified certitude.

Hypothesis 2c: 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. Yet once again, we are unaware of any research specifying or testing these claims.

Empirical Approach

The remainder of this study describes two preregistered²² online survey experiments designed to test the six hypotheses presented in this section. The first experiment examined how respondents evaluated prospective national security decisions based on probability assessments presented in a series of vignettes. The second experiment asked respondents to make their own probability assessments in response to questions regarding foreign policy issues. We describe each of these experiments in separate sections below.

We administered these surveys to a total of 407 national security officials enrolled in two advanced military education programs.²³ Sixty-two percent of these national security officials were active-duty U.S. military officers at the ranks of lieutenant colonel and colonel (for the U.S.

²² See Evidence in Governance and Politics (www.egap.org).

²³ Our larger elite sample comprised students at the National War College, and our smaller elite sample comprised students at Air University. We thank the 407 national security officials in these programs for the time they volunteered to participate in our research.

Army, Air Force, and Marines), or the equivalent U.S. Navy ranks of commander and captain.²⁴ Since the U.S. 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.²⁵ These programs also contained substantial numbers of non-U.S. military officers and civilian officials. Thirteen percent of these respondents were foreign military officers and twenty-five percent were civilians from the U.S. Intelligence Community, Department of State, and other national security-related agencies. We describe respondent demographics for individual experiments below and in supporting material.

We paired these experiments with surveys of 3,017 respondents via Amazon Mechanical Turk (AMT).²⁶ 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, we will show that 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

²⁴ Military institutions distributed our surveys to blocks of students. Response rates exceeded ninety-five percent.

²⁵ 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.

²⁶ We conducted surveys with our National Security Officials and AMT samples between August 5-7, 2015. AMT respondents were required to be U.S. residents at least 18 years of age. We compensated AMT respondents \$2.04 for completing a survey that took an average of 17 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, and Renshon 2015, Williamson 2016.

scholarship that indicates how decision science research conducted on non-elite samples can generate plausible insights into national security decision making.²⁷

Section 2. How Decision Makers Interpret Probability Assessments

We tested Hypotheses 1a-1c by presenting respondents with fictional vignettes involving national security decisions.²⁸ Presented in random order, these scenarios 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.²⁹

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.³⁰ We administered this survey to 208 participants in an advanced military education program, which we call our “National Security

²⁷ See Dhimi et al. 2015 for more discussion of this point.

²⁸ Prominent examples of how international relations scholars have used fictional vignettes to evaluate support for national security decisions include Press, Sagan, and Valentino 2013, Tomz and Weeks 2013, and Kertzer and Brutger 2016.

²⁹ 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 38 below.

³⁰ 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.

Officials Sample.”³¹ We also administered this survey to 1,458 respondents on Amazon Mechanical Turk.³²

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 eighty percent chance (or “very likely”) that the hostages were at the suspected location. This assessment was placed at sixty-five percent (or “likely”) and fifty percent (or “even chance”) for the neutral and pessimistic versions of this scenario, respectively.³³ 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 in order to form their views about proposed actions.

[Figure 2]

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

³¹ This sample was eighty-five percent male, eighty-two percent white, and all respondents possessed a college degree. Eighty-seven percent of respondents were U.S. citizens. Seventy-five percent were military officers.

³² This sample was forty-eight percent male, eighty percent white, with sixty-one percent of respondents possessing a college degree. Average age was thirty-five.

³³ Supporting material shows how probability assessments varied across vignette versions.

refer to this supplementary survey experiment as “Elite Sample B.”³⁴ Altogether, we administered Experiment 1 to 1,857 respondents, who evaluated a total of 5,173 scenarios.

Results

Hypothesis 1a predicts that quantifying probability assessments should make decision makers more likely to support proposed actions. Figure 3 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. The bars in Figure 3 represent the estimated treatment effect associated with expressing these probabilities quantitatively. In supplementary material, we replicate these patterns using multivariate analyses that control for individual attributes.

[Figure 3]

Figure 3 shows that quantifying probability estimates did not make respondents more likely to support proposed actions. All else being equal, respondents were 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.³⁵ Supporting material presents additional analysis of how quantification influenced respondents’ support for proposed actions in all nine possible

³⁴ Eighty-six percent of this sample comprised U.S. citizens; seventy-eight percent were active-duty military. We were asked not to record information on gender and race in this sample.

³⁵ 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.

scenarios (three possible versions of three different vignettes). We found no instance where quantifying probability assessments consistently increased support for proposed actions.

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 Figure 4 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. Moreover, we found no evidence that quantifying probability assessments increased support for proposed actions in optimistic scenarios.³⁶

[Figure 4]

Hypothesis 1c predicts that quantifying probability assessments should make decision makers less willing to delay action in order to gather additional information. Our results refute this hypothesis, as well. Figure 5 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

³⁶ Supplementary material again replicates this finding via multivariate analysis.

decision across respondents was nearly a full point higher in the quantitative assessment condition.³⁷

[Figure 5]

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. This is not necessarily ideal: sometimes, the right move is to act, incomplete information and all.³⁸

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.³⁹ One argument in favor of making probability assessments explicit is that this 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

³⁷ 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), $p=0.001$.

³⁸ 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.

³⁹ Johnson 2004, Kahneman and Renshon 2007, Rapport 2015.

argument, roundly disconfirming three plausible concerns about the drawbacks of probabilistic precision.

Section 3. 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?”).⁴⁰ 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.⁴¹ These surveys produced 61,901 probability estimates.⁴²

We scored qualitative and quantitative estimates using the following procedure (though we note below how our results are robust to several scoring methods). First, we calculated the mean

⁴⁰ 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.

⁴¹ 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 forty-eight percent male and eighty-one percent white. Their average age was thirty-five. Sixty-one percent had a college degree.

⁴² We dropped fourteen estimates because they were greater than one hundred percent, presumably due to typographical errors.

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.⁴³ Using this method, we found that eighty-one percent of AMT respondents and ninety-eight percent of national security officials provided assessments that were more informative, on average, than random guessing.⁴⁴ 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.⁴⁵

Results

Figure 6 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 fourteen

⁴³ 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 is proven false, then the respondent’s Brier Score for that prediction is $(0 - p)^2$.

⁴⁴ Randomly assigning probabilities, with a uniform distribution, to questions with binary outcomes, generates an expected Brier Score of 0.335.

⁴⁵ Tetlock 2005.

percent. Among AMT respondents, the equivalent gap was eleven percent. Both of these comparisons were significant at the $p < 0.001$ level.

[Figure 6]

Alternative scoring rules produced similar patterns. With logarithmic scoring, the difference across treatment conditions in respondents' mean Brier Scores was seventeen percent for National Security Officials and fifteen percent for AMT respondents.⁴⁶ If we round probability estimates to the midpoint of each “word of estimative probability,” instead of using question-specific interpretations as described above, then the gap in performance using Brier Scores was eleven percent for National Security Officials and ten percent for AMT respondents.⁴⁷

Table 1 shows how respondents using numeric probabilities substantially underperformed respondents using words of estimative probability, while controlling for several individual attributes: a 4-point measure of respondent numeracy,⁴⁸ along with indicators for whether respondents were female, U.S. citizens, military officers, and native English speakers.⁴⁹ For AMT respondents we also include a three-point measure of education.⁵⁰ These results show that respondents assessing numeric probabilities obtained consistently higher (that is, worse) Brier Scores.

⁴⁶ 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.

⁴⁷ See supporting information for additional analysis.

⁴⁸ Cokeley et al. 2012.

⁴⁹ In the MTurk sample, the *Military Service* variable indicates respondents with any current or previous military service. Fewer than 1% of MTurk respondents were active-duty military personnel.

⁵⁰ We measured education on a 3-point scale: high school degree or less, college degree (two- or four-year), and postgraduate training. All elite respondents had a college degree and some postgraduate training.

[Table 1]

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 zero to fourteen percent. Yet, as mentioned earlier, 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.⁵¹ First, we scored qualitative estimates according to Mosteller and Youtz’s meta-study of how respondents typically evaluate these terms,⁵² and we rounded numeric estimates to the nearest of these anchors.⁵³ This method found that quantitative assessors produced mean Brier scores that were six percent worse among National Security Officials ($p<0.01$) and four 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 U.S. 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 eight

⁵¹ Again, see supporting information for additional analysis.

⁵² Mosteller and Youtz 1990.

⁵³ 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.”

percent and five percent among National Security Official and AMT respondents, respectively, with both differences retaining statistical significance at the $p < 0.001$ level.⁵⁴ 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.⁵⁵

Precision and certitude

Figure 7 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.⁵⁶ These data suggest support for Hypothesis 2c, that quantifying probability estimates makes analysts more likely to offer assessments with excessive certitude.

[Figure 7]

To evaluate this hypothesis, we define each probability estimate's *Certitude* as the absolute value of its difference from fifty percent.⁵⁷ Table 2 presents ordinary least squares regressions

⁵⁴ See supplementary material for more information.

⁵⁵ Nevertheless, 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.

⁵⁶ All differences in proportions between qualitative and quantitative assessors in Figure 7 are statistically significant at $p < 0.01$. See supporting material for further analysis.

⁵⁷ 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.

examining the accuracy of each probability assessment made by National Security Officials and by AMT respondents. 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.

[Table 2]

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. As shown in supporting material, interaction terms between Quantitative Assessment and these factors were statistically insignificant and 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 five percent among National Security Officials ($p < 0.05$) and to three 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.⁵⁸ This finding raises the possibility that the numbers as a second language problem

⁵⁸ 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 ($p = 0.31$).

mainly appears among respondents with low levels of motivation, and that proper training or more effort would substantially mitigate this bias. We address this possibility in more detail in the study's concluding section.

Section 4. 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 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 which 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.⁵⁹ Similarly, when Mandel and Barnes calibrated a large volume of intelligence estimates, they found analysts' judgments to be *underconfident* on the whole.⁶⁰ 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.⁶¹ Nevertheless, this is a proposition which 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 similar research into other

⁵⁹ Mellers et al. 2014.

⁶⁰ Mandel and Barnes 2014.

⁶¹ 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.

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.

References

- Barnes, Alan. 2016. Making Intelligence Analysis More Intelligent. *Intelligence and National Security* 31 (1): 327-344.
- 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): 351-368.
- Betts, Richard K. 2000. 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.
- Beyth-Marom, Ruth. 1982. How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting* 1: 257-269.
- 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: 508-512.
- 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*. Report Prepared for H.M. Government.

- , David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. Improving Intelligence Analysis with Decision Science. *Perspectives on Psychological Science* 10 (6): 743-757.
- Ellsberg, Daniel. 1961. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics* 75 (4): 643-669.
- Fingar, Thomas. 2011. *Reducing Uncertainty: Intelligence and National Security*. Stanford, Calif.: 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.
- Heuer, Richards J., Jr. 1999. *Psychology of Intelligence Analysis*. Washington, D.C.: 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: 1-12.
- Hyde, Susan D. 2015. Experiments in International Relations: Lab, Survey, and Field. *Annual Review of Political Science* 18: 403-424.
- Jervis, Robert. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, N.J.: Princeton University Press.

- , 2010. *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War*. Ithaca, N.Y: Cornell University Press.
- Johnson, Dominic D. P. 2004. *Overconfidence and War: The Havoc and Glory of Positive Illusions*. Cambridge, Mass.: Harvard University Press.
- Johnston, Rob. 2005. *Analytic Culture in the U.S. Intelligence Community*. Washington, D.C.: 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): 234-249.
- Lanir, Zvi and Daniel Kahneman. 2006. An Experiment in Decision Analysis in Israel in 1975. *Studies in Intelligence* 50 (4).
- Levy, Jack S. 2013. "Psychology and Foreign Policy Decision-Making" in Leonie Huddy, David O. Sears, and Jack S. Levy, eds., *The Oxford Handbook of Political Psychology*, 2nd ed. Oxford, U.K.: Oxford University Press.
- Lowenthal, Mark M. 2006. *Intelligence: From Secrets to Policy*, 3rd ed. Washington, D.C.: CQ Press.
- Mandel, David R. and Alan Barnes. 2014. Accuracy of Forecasts in Strategic Intelligence. *Proceedings of the National Academy of Sciences* 111 (30): 10984-10989.

- 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, Col.: 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: 330-341.
- 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: U.S. Noncombat Operations and Major Wars*. Ithaca, N.Y.: 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): 659-695.

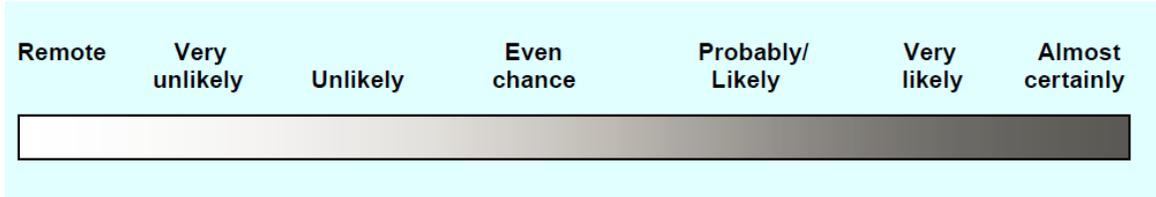
- Rovner, Joshua. 2011. *Fixing the Facts: National Security and the Politics of Intelligence*. Ithaca, N.Y.: Cornell University Press.
- Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: Wiley.
- Shapiro, Jacob N. and Dara Kay Cohen. 2007. Color Blind: Lessons from the Failed Homeland Security Advisory System. *International Security* 32 (2): 121-154.
- Sunstein, Cass R. 2014. *Valuing Life: Humanizing the Regulatory State*. Chicago, IL: University of Chicago Press.
- U.S. Army. 1997. *Field Manual 101-5: Staff Organization and Operations*. Washington, D.C.: Department of the Army.
- U.S. Army. 2009. *Field Manual 5-0: The Operations Process*. Washington, D.C.: Department of the Army.
- Tetlock, Philip E. 2005. *Expert Political Judgment*. Princeton, N.J.: Princeton University Press.
- Tillers, Peter and Jonathan Gottfried. 2006. "Case Comment – *United States v. Copeland*." *Law, Probability, and Risk* 5 (2): 135-157.
- Tomz, Michael R. and Jessica P. Weeks. 2013. Public Opinion and the Democratic Peace. *American Political Science Review* 107 (4): 849-865.
- United States Senate. 2004. *Report of the Select Committee on Intelligence on the U.S. Intelligence Community's Prewar Intelligence Assessments on Iraq*. Washington, D.C.: U.S. Government Printing Office.

Wallsten, Thomas. 1990. "Costs and Benefits of Vague Information" in Robin M. Hogarth, ed., *Insights in Decision Making*. Chicago, Ill.: University of Chicago Press.

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

Zimmer, Alf C. 1984. A Model for the Interpretation of Verbal Predictions. *International Journal of Man-Machine Studies* 20: 121-134.

Guidelines “Explaining Estimative Language” in National Intelligence Estimates (2007)



U.S. 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%

Defense Intelligence Agency memorandum on “Expressing Analytic Certainty” (2015)

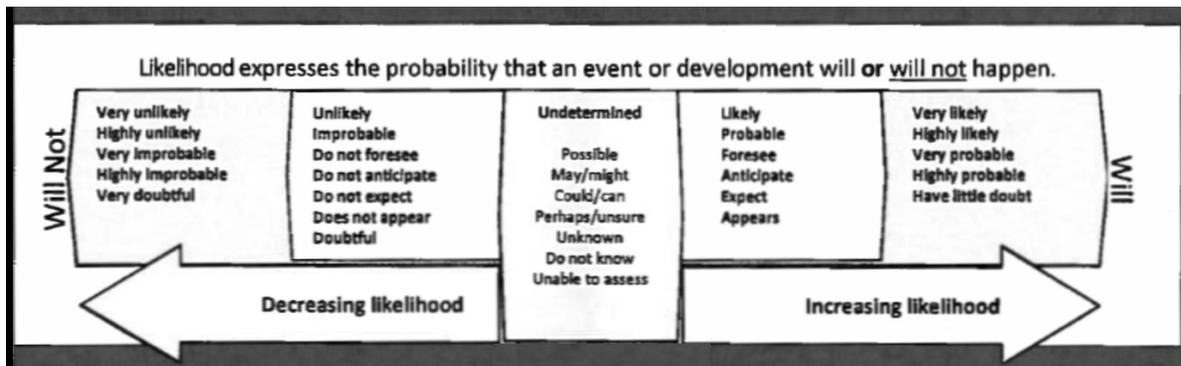


Figure 1. Three guidelines for expressing probability in intelligence

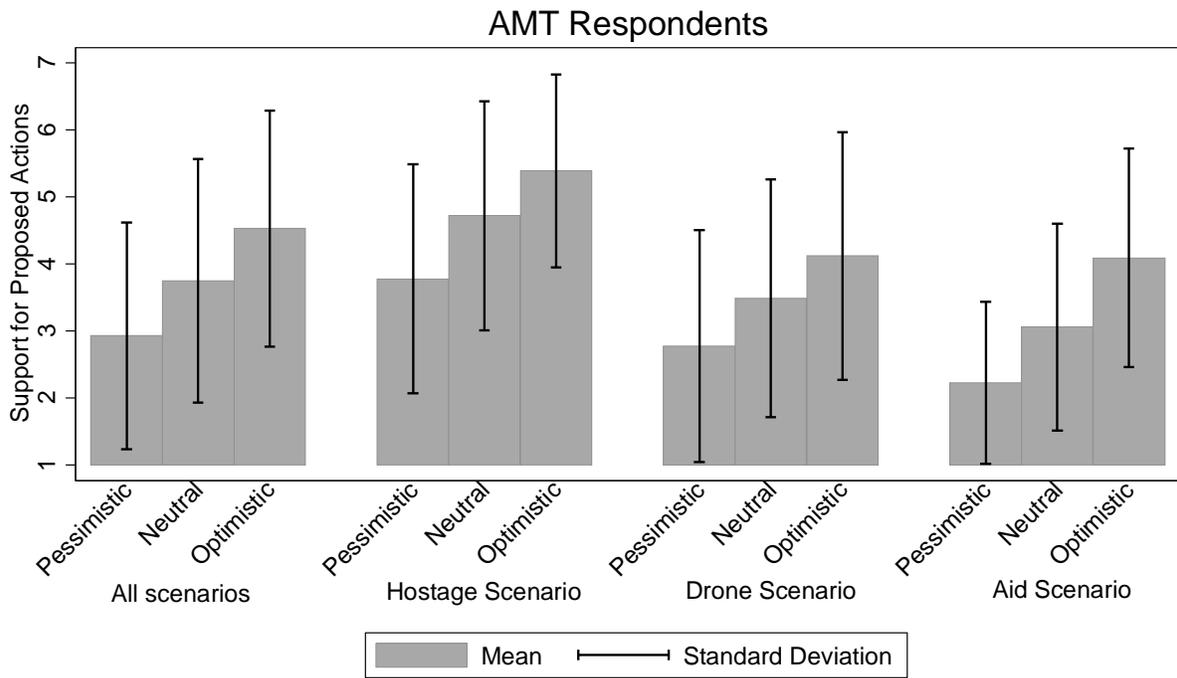
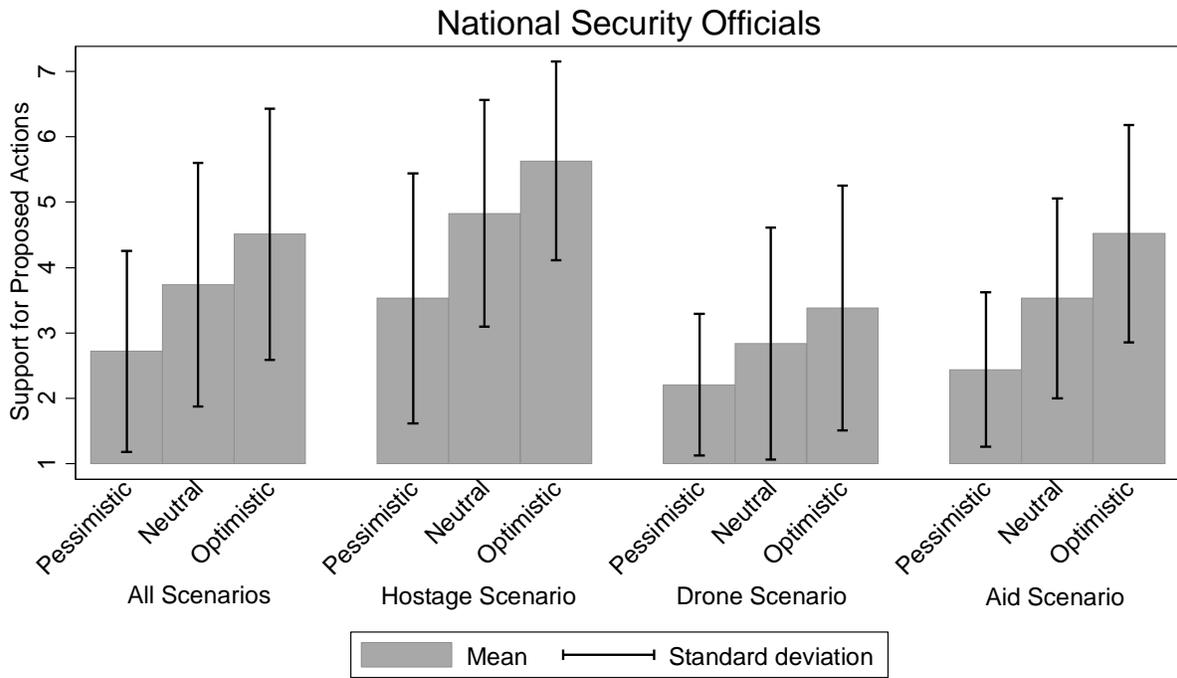


Figure 2. Support for proposed actions across scenarios

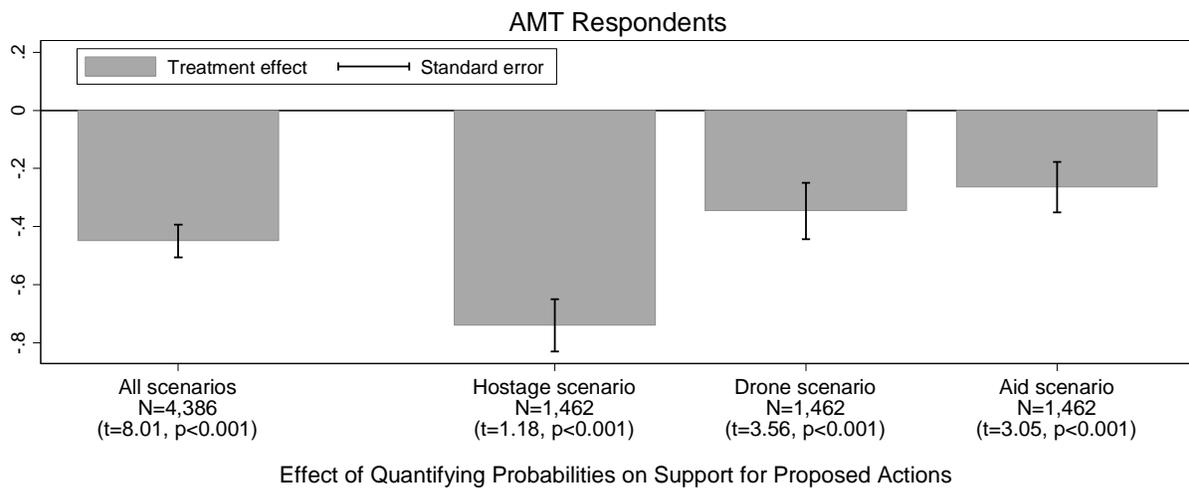
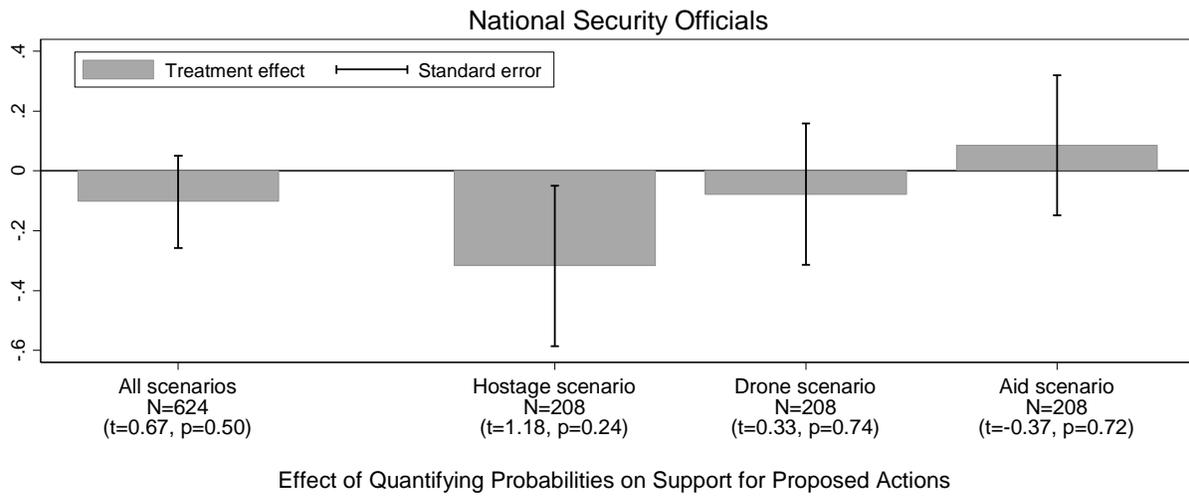


Figure 3. How quantifying probabilities influenced support for proposed actions

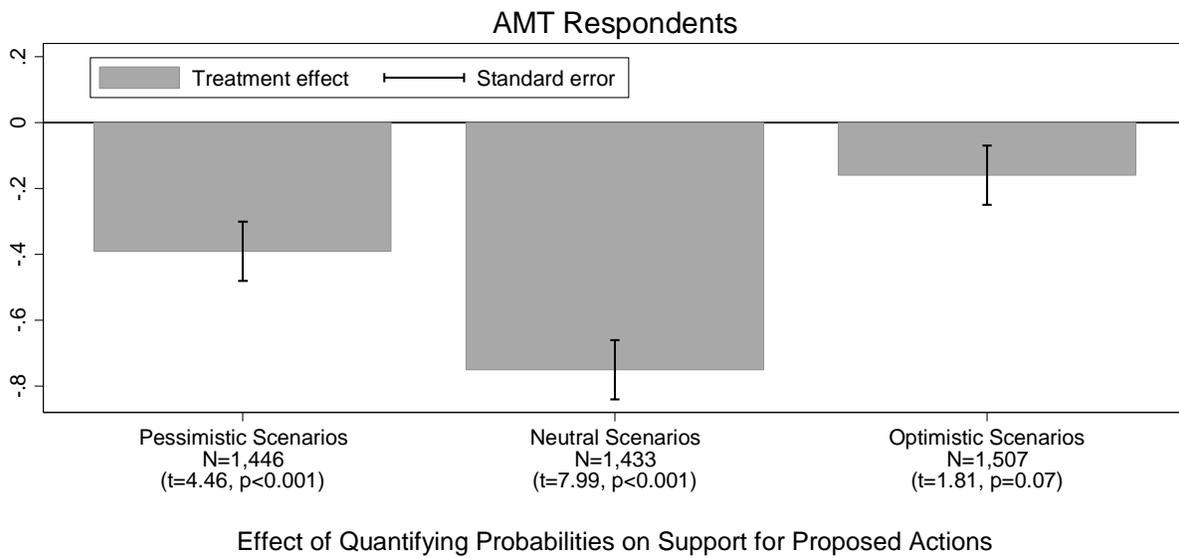
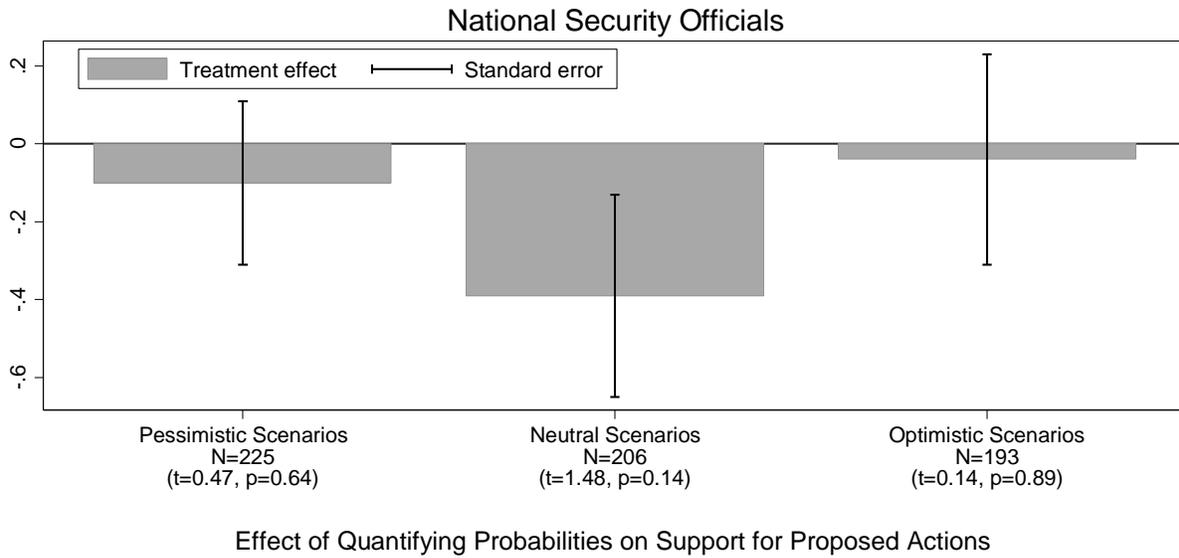


Figure 4. Impact of quantification on support for proposed action across scenario types

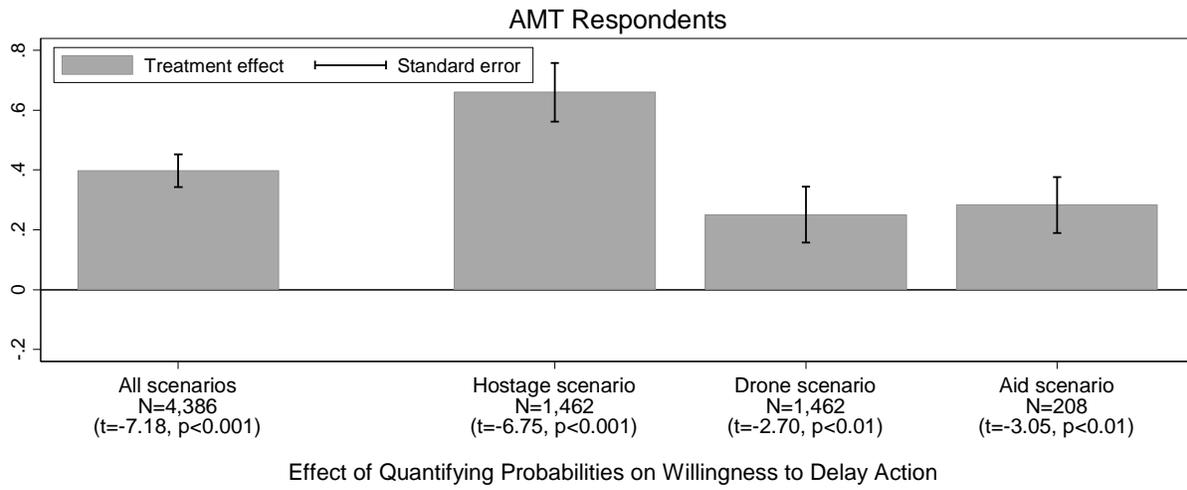
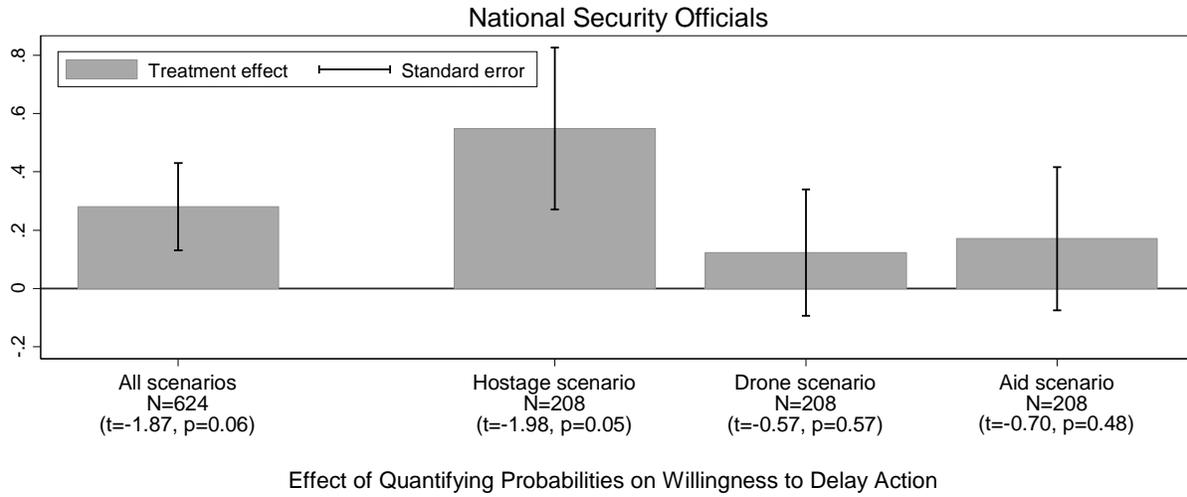


Figure 5. How quantifying probabilities influenced willingness to delay action

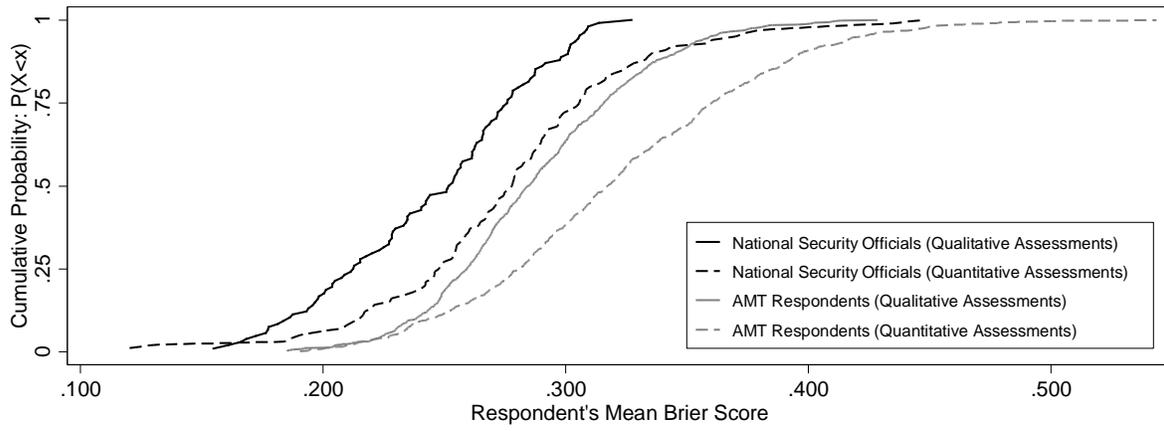


Figure 6. Cumulative distributions of respondent Brier scores by treatment group

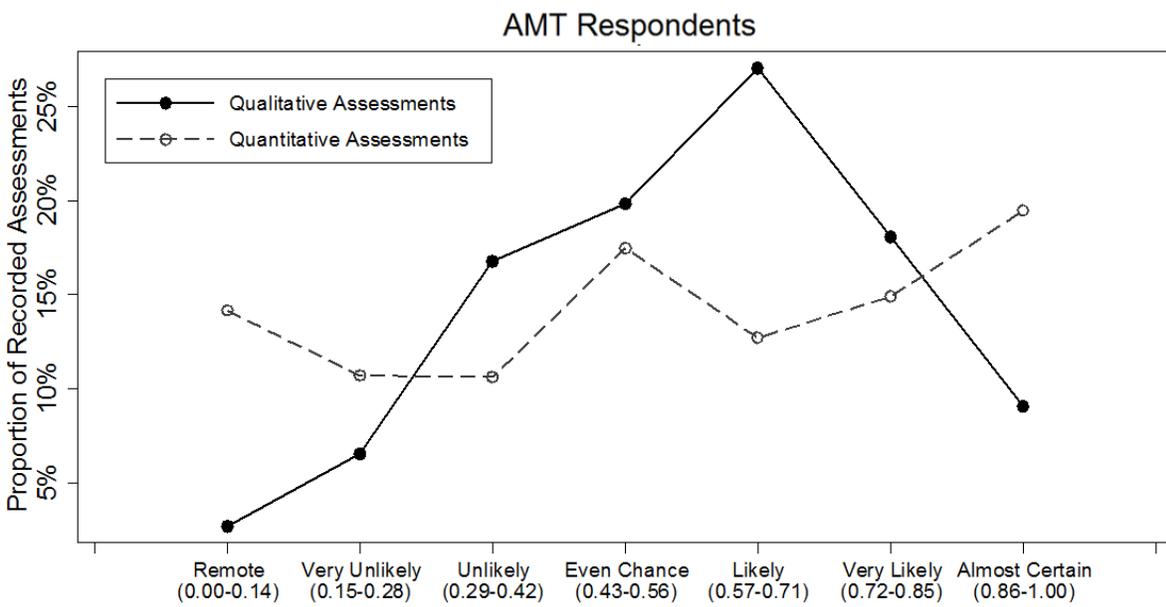
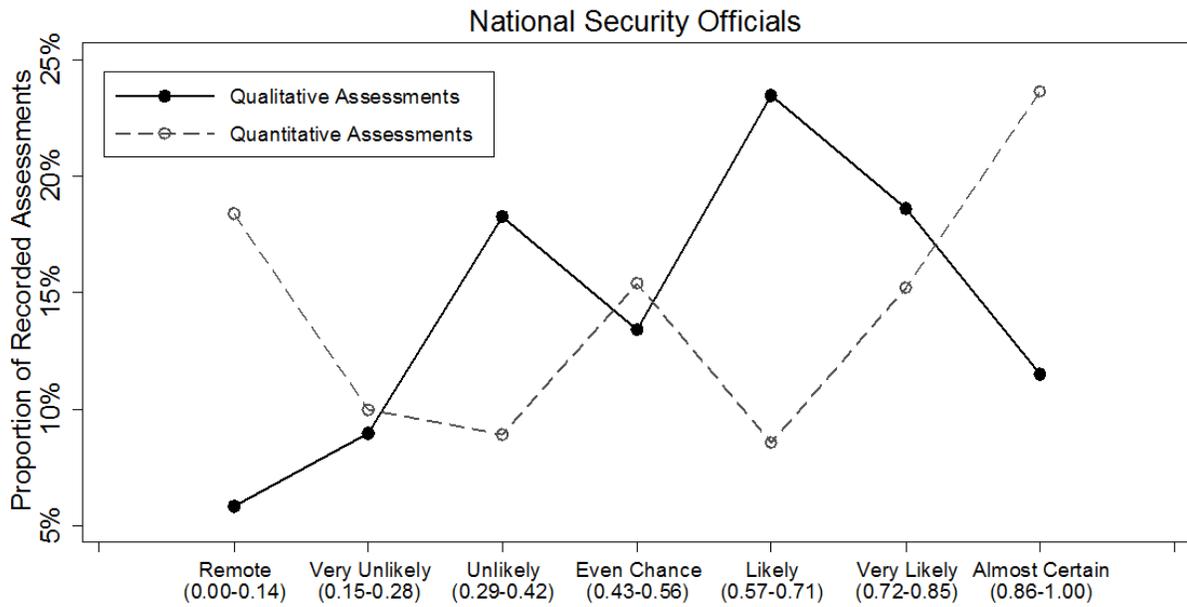


Figure 7. Comparing distributions of qualitative/quantitative probability assessments

	National Security Officials	AMT Respondents
<i>Quantitative Assessment</i>	0.034 (.01) ^{***}	0.033 (2.4e ⁻³) ^{***}
<i>Numeracy</i>	-0.003 (2.5e ⁻³)	-0.004 (1.0e ⁻³) ^{***}
<i>Military Service</i>	-0.001 (.01)	0.025 (.03)
<i>Female</i>	0.003 (.01)	0.004 (2.4e ⁻³)
<i>U.S. Citizen</i>	0.014 (.01)	-7.40e ⁻⁴ (.01)
<i>English Native Lang.</i>	-0.022 (.01)	0.005 (.01)
<i>Education Level</i>		-0.010 (1.7e ⁻³) ^{***}
<i>Constant</i>	0.244 (.01) ^{***}	0.296 (.01) ^{***}
N	208 respondents	1,561 respondents
R ²	0.184	0.153

Ordinary least squares regression predicting respondents' mean Brier Scores. Lower Brier Scores indicate more accurate assessments. * p<0.05, ** p<0.01, *** p<0.001. Robust standard errors.

Table 1. Predictors of Brier Scores

	<i>National Security Officials</i>		<i>AMT respondents</i>	
	Model 1	Model 2	Model 3	Model 4
<i>Quantitative Assessment</i>	0.032 (.01) ^{***}	0.001 (.01)	0.013 (2.1e ⁻³) ^{***}	-0.002 (2.4e ⁻³)
<i>Certitude</i>	0.032 (.02)	-0.040 (.03)	0.297 (.01) ^{***}	0.255 (.01) ^{***}
<i>Quantitative Assessment x Certitude</i>		0.124 (.04) ^{***}		0.070 (.02) ^{***}
<i>Forecast</i>	0.223 (.02) ^{***}	0.222 (.02) ^{***}	-0.237 (.01) ^{***}	-0.237 (.01) ^{***}
<i>Numeracy</i>	-0.003 (2.5e ⁻³)	-0.003 (2.5e ⁻³)	-0.004 (9.7e ⁻⁴) ^{***}	-0.004 (9.6e ⁻⁴) ^{***}
<i>Military Service</i>	-0.001 (.01)	-0.001 (.01)	0.015 (.02)	0.013 (.02)
<i>Female</i>	0.003 (.01)	0.003 (.01)	0.009 (2.2e ⁻³) ^{***}	0.009 (2.2e ⁻³) ^{***}
<i>U.S. Citizen</i>	0.011 (.01)	0.015 (.01)	0.003 (.01)	0.002 (.01)
<i>English Native Lang.</i>	-0.020 (.01)	-0.022 (.01)	0.003 (.01)	0.003 (.01)
<i>Education Level</i>			-0.011 (1.6e ⁻³) ^{***}	-0.011 (1.6e ⁻³) ^{***}
<i>Constant</i>	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

Ordinary least squares regression predicting Brier scores for individual probability assessments. Robust standard errors clustered by respondent. Question-level fixed effects not shown. Note that lower Brier Scores indicate more accurate assessments. *p<.05, **p<.01, ***p<.001.

Table 2. Relationship between quantitative assessment, respondent certitude, and accuracy

Appendix. Scenario Text

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 U.S. 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. U.S. 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 U.S. 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 U.S. 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*

Remote Very unlikely Unlikely Even chance Probably/Likely Very likely Almost certainly

Probability estimates can range from "remote" to "almost certainly"

Hostage Scenario, Neutral Version, Quantitative Assessments

ISIS is holding three American aid workers hostage. The U.S. 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. U.S. 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 U.S. 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 U.S. 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*

Supporting Material for “Behavioral Consequences of Probabilistic Precision”

This supplement contains the following material:

§A. Respondent demographics

Supplementary analysis of Survey Experiment 1: How decision makers interpret probabilities

§B. Full text of drone and aid scenarios

§C. Response measure wordings

§D. Variation in probability assessments across scenario versions

§E. Analysis of each scenario version

§F. T-tests for Elite Sample B

§G. Multivariate analysis of scenario data

Supplementary analysis of Survey Experiment 2: How analysis estimate probabilities

§H. Question list for Survey Experiment 2

§I. Distributions of assessments across question types

§J. Additional rescoring of probability assessments

§K. Interactions between quantitative assessments and respondent attributes

§L. Replication of Survey Experiment 2 results using logarithmic scoring

§A. Respondent demographics

The paper presents survey experiment results from three samples of respondents. National Security Officials took both Survey Experiment 1 and Survey Experiment 2 (in random order). Amazon Mechanical Turk (AMT) respondents were randomly assigned to take only one of the two surveys. We report demographics for both AMT populations separately. Elite Sample B responded to an abridged version of Survey Experiment 1, containing only the neutral hostage rescue scenario. Due to concerns for protecting respondent anonymity, we were able to ask fewer demographic questions of elite sample respondents than of AMT respondents.

The Berlin adaptive numeracy test (Cokely et al. 2012) divides respondents into four categories based on their responses to 2-3 word problems (the choice of problems depends on a respondent's answers). Previous research with nationally-representative samples generally divides responses into four, equally-sized categories, in which higher scores indicate higher numeracy. We report the percentage of respondents falling into each of these categories below.

Table S1. Respondent demographics

	National Security Officials (N=208)	AMT Respondents, Survey 1 (N=1,458)	AMT Respondents, Survey 2 (N=1,561)	Elite Sample B (N=199)
% Female	15%	52%	52%	-
% White	82%	80%	81%	-
% College degree	100%	61%	61%	100%
% U.S. citizen	87%	99%	99%	86%
% English as first language	85%	98%	98%	86%
% Current military service	75%	0.8%	0.4%	78%
% Current or former mil. service	84%	6%	5%	-
Age	-	34.7 (11.1)	34.5 (11.3)	-
% Numeracy-1	29%	40%	42%	-
% Numeracy-2	29%	21%	19%	-
% Numeracy-3	11%	17%	18%	-
% Numeracy-4	31%	22%	21%	-

§B. Full text of drone and aid scenarios

The manuscript's appendix presented the full text of the hostage rescue scenario that we presented to respondents in our first survey experiment. Here, we present full text for the other two vignettes, involving a drone strike and a decision to support local security forces in counterinsurgency.

Drone Scenario, Neutral Version, Qualitative Assessment

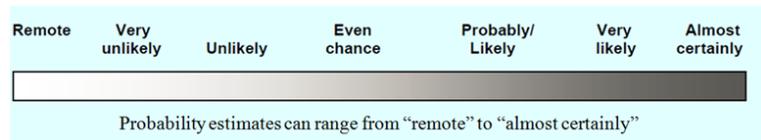
The Central Intelligence Agency uses drones to monitor houses in Yemen believed to be used by Al Qaeda in the Arabian Peninsula (AQAP). CIA analysts report that an unusual number of people have recently been gathering at one of these houses. At any given time, there are at least 8-12 individuals inside the house. All of these individuals appear to be male, but it is impossible to confirm their identities. Recent intercepted communications have indicated that AQAP's senior leadership was planning to convene in this area.

Intelligence analysts stress that their judgments are subjective and that they are based on incomplete information. However, based on all available intelligence, analysts assess that it is likely that the house contains members of AQAP's senior leadership. Drone operators are standing by to attack the house. They believe it is very likely that a drone strike on the house would kill everyone inside.

Analysts warn that it is possible, though unlikely, that the house contains women and children. If U.S. forces strike this target, then it is almost certain that AQAP would not meet again anywhere in this region. This would compromise ongoing surveillance efforts in the area. It is not clear when U.S. intelligence will have another lead like this one.

Summary of estimated chances:

- The house contains members of Al Qaeda's senior leadership: *Likely*
- A drone strike on the house would kill everyone inside: *Very likely*
- The house contains women and children: *Unlikely*
- The drone strike will compromise ongoing surveillance efforts in the area: *Almost certainly*



Security Forces Scenario, Neutral Version, Qualitative Assessment

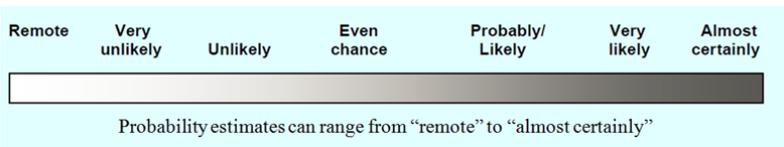
An Afghan leader named Ghamay Jan recently approached U.S. officials. Jan offered to mobilize 500 followers to combat the Taliban along a dangerous stretch of the border with Pakistan located in Khost Province. Jan requests that the United States provide him funding, equipment, and permission to use force against the Taliban.

Intelligence analysts stress that their judgments are subjective and that they are based on incomplete information. Nevertheless, they believe it is likely that Jan can mobilize the forces he has promised. Moreover, Jan’s followers have substantial military experience and extensive family ties in Khost Province. If they cooperate with the United States, analysts believe it is likely that they would prevent the Taliban from crossing the nearby border.

Yet Ghamay Jan is a controversial figure. Intelligence analysts believe there is an even chance that he previously assisted the Taliban to establish a presence in this part of Khost Province. They also say it is very likely that Jan would use the authority he requests to facilitate illegal smuggling. If the United States supports Ghamay Jan, analysts say it is unlikely that they can retain the backing of other local leaders in Khost Province. Yet those leaders have been unable to secure their border with Pakistan in the past.

Summary of estimated chances:

- Jan can mobilize the forces he has promised: *Likely*
- If Jan’s forces cooperate with the United States, they would prevent Taliban from crossing the nearby border: *Likely*
- Jan previously assisted the Taliban: *Even chance*
- Jan will use the authority he requests to secure illegal smuggling: *Very likely*
- The United States can retain backing from other local leaders if U.S. forces support Ghamay Jan: *Unlikely*



§C. Response measure wordings

Following each vignette, respondents were presented with the following questions.

On the following scale of 1-7, please rate your level of support for [approving the hostage rescue mission immediately / approving the proposed drone strike immediately / backing Ghamay Jan]

1: Strongly oppose 2 3 4: Neither support nor oppose 5 6 7: Strongly support

On the following scale of 1-7, please rate your level of support for [delaying decision 1-2 weeks to gather additional information about the compound / waiting another 1-2 days to gather more information about who might be meeting in this location / waiting another 6 months to gather more information about whether backing Ghamay Jan is a better option than backing local officials]

1: Strongly oppose 2 3 4: Neither support nor oppose 5 6 7: Strongly support

On the following scale of 1-7, please rate your level of confidence in making these assessments

1: No confidence at all 2 3 4: Moderate confidence 5 6 7: Complete confidence

§D. Variation in probability assessments across scenario versions

As described in the main text, we presented each scenario in one of three versions. These versions – which we label “pessimistic,” “neutral,” and “optimistic,” below – involved different probability assessments intended to shape respondents’ views about supporting risky actions. As shown in the main text (Table 1), these variations systematically influenced respondents’ views. Table S2 shows how we varied probability assessments across scenario versions.

Table S2. Variations in Probability Assessments Across Scenario Versions

<i>Assessment</i>	<i>Probability Assessments Across Scenario Versions</i>		
	<i>Pessimistic</i>	<i>Neutral</i>	<i>Optimistic</i>
<i>Hostage Rescue Scenario</i>			
Hostages at compound	Even Chance (50 percent)	Likely (65 percent)	Very Likely (80 percent)
Special forces can retrieve hostages	Likely (65 percent)	Very Likely (80 percent)	Almost Certain (95 percent)
Soldiers wounded on mission	Likely (65 percent)	Even Chance (50 percent)	Unlikely (35 percent)
Collateral damage	Even Chance (50 percent)	Unlikely (35 percent)	Very Unlikely (20 percent)
Hostages killed if mission fails	Almost Certain (95 percent)	Almost Certain (95 percent)	Almost Certain (95 percent)
<i>Drone Strike Scenario</i>			
House contains Qaeda leaders	Even Chance (50 percent)	Likely (65 percent)	Very Likely (80 percent)
Drone strike kills occupants	Very Likely (80 percent)	Very likely (80 percent)	Very Likely (80 percent)
House contains women/children	Likely (65 percent)	Unlikely (35 percent)	Remote Chance (5 percent)
Strike compromises surveillance	Almost Certain (95 percent)	Almost Certain (95 percent)	Almost Certain (95 percent)
<i>Local Security Forces Scenario</i>			
Jan can mobilize forces	Even Chance (50 percent)	Likely (65 percent)	Very Likely (80 percent)
Jan's forces can secure border	Unlikely (35 percent)	Likely (65 percent)	Almost Certain (95 percent)
Jan previously assisted Taliban	Very likely (80 percent)	Even Chance (50 percent)	Very unlikely (20 percent)
Jan will secure illegal smuggling	Almost Certain (95 percent)	Very Likely (80 percent)	Likely (65 percent)
U.S. can retain local leaders' support	Very Unlikely (20 percent)	Unlikely (35 percent)	Even Chance (50 percent)

§E. Analysis of scenario versions

Survey Experiment 1 involved three different scenarios: a hostage rescue mission, a drone strike, and a proposal to sponsor local security forces. We presented each scenario in three different versions: pessimistic, neutral, and optimistic. Table S3 describes support for risky actions within each scenario version. The key finding from this analysis is that there is no instance in which quantifying probability assessments increases support for risky action in a statistically significant fashion.

Table S3. Scenario Version Responses

Scenario	Treatment Condition	Scenario version		
		Pessimistic	Neutral	Optimistic
<u>National Security Officials</u>				
All	<i>Qualitative probabilities</i>	2.84 (1.68)	3.97 (1.88)	4.54 (1.83)
	<i>Quantitative probabilities</i>	2.74 (1.48)	3.58 (1.87)	4.50 (1.96)
		$p=0.641$	$p=0.139$	$p=0.886$
Hostage	<i>Qualitative probabilities</i>	3.79 (1.96)	5.26 (1.57)	5.70 (1.41)
	<i>Quantitative probabilities</i>	3.35 (1.94)	4.34 (1.86)	5.51 (1.59)
		$p=0.346$	$p=0.028$	$p=0.622$
Drone	<i>Qualitative probabilities</i>	2.24 (1.26)	3.21 (1.89)	3.32 (1.72)
	<i>Quantitative probabilities</i>	2.37 (1.10)	2.66 (1.76)	3.48 (2.02)
		$p=0.630$	$p=0.227$	$p=0.731$
Security	<i>Qualitative probabilities</i>	2.38 (1.18)	3.47 (1.48)	4.74 (1.54)
	<i>Quantitative probabilities</i>	2.61 (1.23)	3.67 (1.54)	4.38 (1.74)
		$p=0.418$	$p=0.611$	$p=0.384$
<u>AMT Respondents</u>				
All	<i>Qualitative probabilities</i>	3.12 (1.72)	4.13 (1.83)	4.61 (1.77)
	<i>Quantitative probabilities</i>	2.73 (1.64)	3.37 (1.72)	4.44 (1.74)
		$p=0.000$	$p=0.00$	$p=0.071$
Hostage	<i>Qualitative probabilities</i>	4.18 (1.59)	5.20 (1.57)	5.52 (1.38)
	<i>Quantitative probabilities</i>	3.40 (1.74)	4.22 (1.71)	5.23 (1.50)
		$p=0.000$	$p=0.000$	$p=0.023$
Drone	<i>Qualitative probabilities</i>	2.97 (1.77)	3.90 (1.75)	4.03 (1.87)
	<i>Quantitative probabilities</i>	2.59 (1.67)	3.04 (1.68)	4.21 (1.82)
		$p=0.016$	$p=0.000$	$p=0.273$
Security	<i>Qualitative probabilities</i>	2.30 (1.22)	3.28 (1.63)	4.22 (1.64)
	<i>Quantitative probabilities</i>	2.16 (1.20)	2.84 (1.42)	3.95 (1.62)
		$p=0.221$	$p=0.001$	$p=0.068$

§F. T-tests for Elite Sample B

Table S4 presents two-way t-tests analyzing responses to the neutral hostage scenario for Elite Sample B. These results support the analyses presented in Section 2: quantifying probability assessments reduces support for risky action, increases support for gathering additional information, and had no significant impact on respondents' confidence levels.

Table S4. Survey results from Elite Sample B

	Support for hostage rescue (1-7 scale)	Support for delaying decision (1-7 scale)	Confidence in assessment (1-7 scale)
<i>Qualitative assessments</i>	5.33 (1.56)	3.14 (1.97)	5.18 (1.19)
<i>Quantitative assessments</i>	4.53 (1.86)	4.11 (2.07)	5.13 (1.26)
	$p=0.001$	$p=0.001$	$p=0.793$

§G. Ordered probit analysis of scenario data

The manuscript analyzes responses to Survey Experiment 1 using simple comparisons of means. Here, we replicate those findings using multivariate analysis. Since the dependent variables in this analysis are each ordered on a 1-7 scale, we replicate that analysis below, in Tables S5a/S5b, using ordered probit regressions. Here, we use interaction terms to test the proposition that quantifying probability assessments amplified cues in the *Optimistic* and *Pessimistic* vignettes.

Note that, while these analyses show a positive interaction terms between quantitative assessment and optimistic scenarios, those coefficients are smaller than what we estimate for the *Quantitative Assessment* indicator itself. This refutes the expectation that quantifying probability assessments would make respondents more likely to support proposed actions in optimistic scenarios. Note also how that interaction terms between quantitative assessment and pessimistic scenarios are positive, which refutes the expectation that quantifying probability assessments will depress support for proposed actions in the pessimistic scenarios.

Table S5a. Responses to Scenarios – National Security Officials (ordered probit)

	<i>Model 1:</i> Predicting support for risky action	<i>Model 2:</i> Predicting support for risky action, with interaction terms	<i>Model 3:</i> Predicting support for delaying action
<i>Quantitative assessment</i>	-0.099 (.09)	-0.254 (.15)	0.186 (.09)*
<i>Optimistic scenario</i>	0.488 (.11)***	0.418 (.15)**	-0.361 (.11)***
<i>Pessimistic scenario</i>	-0.559 (.10)***	-0.692 (.13)***	0.243 (.10)*
<i>Hostage scenario</i>	0.738 (.10)***	0.743 (.10)***	-0.140 (.10)
<i>Drone scenario</i>	-0.368 (.10)***	-0.370 (.10)***	0.698 (.11)***
<i>Numeracy</i>	-0.110 (.04)**	-0.108 (.04)**	0.080 (.04)*
<i>Female</i>	-0.186 (.12)	-0.188 (.13)	0.103 (.14)
<i>Military officer</i>	0.122 (.10)	0.120 (.10)	-0.085 (.11)
<i>U.S. citizen</i>	0.110 (.28)	0.096 (.28)	0.028 (.31)
<i>English as native lang.</i>	0.071 (.28)	0.072 (.28)	-0.128 (.32)
<i>Optimistic scenario</i> <i>x Quantitative assessment</i>		0.159 (.22)	
<i>Pessimistic scenario</i> <i>x Quantitative assessment</i>		0.292 (.20)	
<i>Cut-points</i>	-1.339, -0.422, 0.030, 0.266, 0.839, 1.783	-1.418, -0.499, -0.046, 0.190, 0.763, 1.709	-1.513, -0.870, -0.353, -0.032, 0.372, 0.999
Pseudo R ²	0.087	0.088	0.048

Ordered probit regressions predicting 7-point response measures with robust standard errors.

* p<0.05 ** p<0.01 *** p<0.001. All models have 624 observations.

Table S5b. Responses to Scenarios – AMT Respondents (ordered probit)

	<i>Model 1:</i> Predicting support for risky action	<i>Model 2:</i> Predicting support for risky action, with interaction terms	<i>Model 3:</i> Predicting support for delaying action
<i>Quantitative assessment</i>	-0.261 (.03)***	-0.455 (.05)***	0.227 (.03)***
<i>Optimistic scenario</i>	0.460 (.04)***	0.276 (.05)***	-0.271 (.04)***
<i>Pessimistic scenario</i>	-0.522 (.04)***	-0.613 (.05)***	0.137 (.04)***
<i>Hostage scenario</i>	0.939 (.04)***	0.945 (.04)***	-0.172 (.04)***
<i>Drone scenario</i>	0.208 (.04)***	0.208 (.04)***	0.306 (.04)***
<i>Numeracy</i>	-0.074 (.01)***	-0.073 (.01)***	0.033 (.01)*
<i>Female</i>	-0.000 (.03)	-0.005 (.03)	0.150 (.03)***
<i>Military service</i>	0.082 (.08)	0.079 (.08)	-0.205 (.08)**
<i>U.S. citizen</i>	0.041 (.15)	0.035 (.15)	-0.214 (.16)
<i>English as native lang.</i>	-0.131 (.11)	-0.128 (.11)	-0.265 (.14)
<i>Education</i>	0.048 (.02)*	0.048 (.02)*	-0.009 (.02)
<i>Age</i>	8.7e ⁻⁴ (1.5e ⁻³)	9.4e ⁻⁴ (1.5e ⁻³)	2.6e ⁻³ (1.5e ⁻³)
<i>Optimistic scenario</i> <i>x Quantitative assessment</i>		0.386 (.08)***	
<i>Pessimistic scenario</i> <i>x Quantitative assessment</i>		0.183 (.08)*	
<i>Cut-points</i>	-1.154, -0.486, 0.087, 0.398, 1.056, 1.737	-1.252, -0.582, -0.007 0.305, 0.966, 1.648	-1.782, -1.347, -0.929, -0.548, -0.093, 0.463
Pseudo R ²	0.074	0.076	0.022

Ordinary least squares regressions predicting 7-point response measures with robust standard errors.

* p<0.05 ** p<0.01 *** p<0.001. All models have 4,376 observations.

§H. Question list for Survey Experiment 2

Table S6 presents the question list for survey experiment 2. Thirty of those questions had known answers as of the date these surveys were administered (August 5-7, 2015). Five questions, labeled “unknown,” had answers that were unknowable at the time. Five questions, labeled “forecast,” involve predictions, were evaluated on February 6, 2016.

As described in the paper, we varied question types in this way so that we could assess whether or not any biases we identified in our analysis were confined to a particular kind of estimate. As shown in part F of this supplement, below, we found that respondents who provided numeric estimates employed noticeably greater certitude than respondents who used words of estimative probability across all three question types.

Table S6. Question List for Survey Experiment 2

All assessments recorded between August 5-7, 2015; outcomes are coded relative to that date.

	<i>Question Text</i>	<i>Outcome</i>
Q1	In your opinion, what are the chances that Afghanistan's literacy rate is currently above 50 percent?	0
Q2	In your opinion, what are the chances that Saudi Arabia currently exports more oil than all other countries in the world combined?	0
Q3	In your opinion, what are the chances that the United States currently has a longer life expectancy than Jamaica?	1
Q4	In your opinion, what are the chances that the United States currently operates a military base in Ethiopia?	0
Q5	In your opinion, what are the chances that the United States has an active territorial claim in Antarctica?	0
Q6	In your opinion, what are the chances that France currently has more soldiers stationed in Afghanistan than any NATO member besides the United States?	0
Q7	In your opinion, what are the chances that more than 20 countries currently operate nuclear power plants?	1
Q8	In your opinion, what are the chances that Japan is currently a member of the International Whaling Commission?	1
Q9	In your opinion, what are the chances that Russia is a member of the Nuclear Nonproliferation Treaty?	1
Q10	In your opinion, what are the chances that the United States currently has free trade agreements in place with fewer than 30 countries?	1

Q11	In your opinion, what are the chances that fewer than 80 countries currently recognize Taiwan's independence from China?	1
Q12	In your opinion, what are the chances that ISIS draws more foreign fighters from Egypt than from any other country outside of Iraq and Syria?	0
Q13	In your opinion, what are the chances that Russia's economy grew in 2014?	0
Q14	In your opinion, what are the chances that Haiti has the lowest per capita income of any Latin American country?	1
Q15	In your opinion, what are the chances that there are currently more Muslims in the world than there are Roman Catholics?	1
Q16	In your opinion, what are the chances that Sweden is a member of NATO?	0
Q17	In your opinion, what are the chances that Tokyo's stock exchange is the second largest stock exchange in the world?	1
Q18	In your opinion, what are the chances that the U.S. State Department currently lists Iran as a state sponsor of terrorism?	1
Q19	In your opinion, what are the chances that the Arabic media organization al-Jazeera currently operates bureaus in more countries than does CNN?	1
Q20	In your opinion, what are the chances that the United States currently possesses more than 2,000 nuclear warheads?	1
Q21	In your opinion, what are the chances that the economy of North Korea is larger than the economy of New Hampshire?	0
Q22	In your opinion, what are the chances that German President Angela Merkel is currently the longest-serving head of government in Western Europe?	1
Q23	In your opinion, what are the chances that there are currently more refugees living in Lebanon than in any other country in the world?	0
Q24	In your opinion, what are the chances that the United States currently conducts more trade with Mexico than with the European Union?	1
Q25	In your opinion, what are the chances that the largest U.S. Embassy is currently located in Beijing?	0
Q26	In your opinion, what are the chances that the U.S. defense budget is more than five times as large as China's defense budget?	0
Q27	In your opinion, what are the chances that the United States currently operates more aircraft carriers than all other countries in the world combined?	0
Q28	In your opinion, what are the chances that Israel receives more foreign aid than any other country in the world?	0
Q29	In your opinion, what are the chances that more than 3 million people live within the borders of the Palestinian territories of the West Bank and Gaza?	1
Q30	In your opinion, what are the chances that more than 5,000 people died as a result of the Ebola outbreak in West Africa in 2014?	1
Q31	In your opinion, what are the chances that Ukrainian rebels knew that Malaysia Airlines Flight 17 was a civilian aircraft before they shot it down?	Unknown
Q32	In your opinion, what are the chances that Iran's Supreme Leader, Ayatollah Khamenei, currently intends to develop a nuclear weapon?	Unknown
Q33	In your opinion, what are the chances that the United States would have invaded Iraq if the September 11 terrorist attacks had not occurred?	Unknown
Q34	In your opinion, what are the chances that without waterboarding captured terrorists, there would have been at least one more major terrorist attack (>1,000 casualties) on U.S. soil since 2001?	Unknown
Q35	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?	Unknown
Q36	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?	Forecast: 0

- Q37 In your opinion, what are the chances that **within the next six months, the Iraqi Security Forces will reclaim control of either Ramadi or Mosul (or both) from ISIS?** Forecast: 1
- Q38 In your opinion, what are the chances that **there will be a new Pope within the next six months?** Forecast: 0
- Q39 In your opinion, what are the chances that **more than 50,000 U.S. citizens will travel to Cuba within the next six months?** Forecast: 1
- Q40 In your opinion, what are the chances that **more than 10 U.S. soldiers will be killed fighting ISIS within the next six months?** Forecast: 0

§I. Distributions of assessments across question types

The paper demonstrated that respondents made probability assessments with noticeably greater certitude when using quantitative as opposed to qualitative expressions. Those data only included questions with knowable answers (i.e., the questions for which we scored respondents' performance).

Tables S7a and S7b show how the same pattern holds when we examine probability assessments that respondents made in response to forecasts or to questions with answers that were unknowable at the time. Statistical significance estimated via two-way t-tests.

Table S7a. Proportion of estimates registered in different segments of the number line (National Security Officials)

	Remote (0.00- 0.14)	Very Unlikely (0.15- 0.28)	Unlikely (0.29- 0.42)	Even Chance (0.43- 0.56)	Likely (0.57- 0.71)	Very Likely (0.72- 0.85)	Almost Certain (0.86- 1.00)
<i>Questions with knowable answers</i>							
<i>Qualitative assessments</i>	4.51	7.10	16.67	12.87	25.28	20.34	13.24
<i>Quantitative assessments</i>	15.73	8.73	8.60	15.60	8.87	16.33	26.13
	<i>p=0.00</i>	<i>p=0.02</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>
<i>Questions about unknown states of the world</i>							
<i>Qualitative assessments</i>	7.59	11.67	19.26	14.44	19.63	17.96	9.44
<i>Quantitative assessments</i>	20.20	11.40	8.00	13.80	9.00	14.20	23.40
	<i>p=0.00</i>	<i>p=0.89</i>	<i>p=0.00</i>	<i>p=0.77</i>	<i>p=0.10</i>	<i>p=0.00</i>	<i>p=0.00</i>
<i>Forecasts</i>							
<i>Qualitative assessments</i>	12.04	17.41	26.67	15.56	16.30	8.89	3.15
<i>Quantitative assessments</i>	32.40	16.00	11.60	15.60	6.20	9.40	8.80
	<i>p=0.00</i>	<i>p=0.54</i>	<i>p=0.00</i>	<i>p=0.98</i>	<i>p=0.78</i>	<i>p=0.00</i>	<i>p=0.00</i>

Table S7b. Proportion of estimates registered in different segments of the number line (AMT Respondents)

	Remote (0.00- 0.14)	Very Unlikely (0.15- 0.28)	Unlikely (0.29- 0.42)	Even Chance (0.43- 0.56)	Likely (0.57- 0.71)	Very Likely (0.72- 0.85)	Almost Certain (0.86- 1.00)
	<i>Questions with knowable answers</i>						
<i>Qualitative assessments</i>	1.65	4.94	15.66	19.75	29.13	19.30	9.57
<i>Quantitative assessments</i>	10.90	9.98	10.64	18.12	13.66	16.12	20.58
	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>
	<i>Questions about unknown states of the world</i>						
<i>Qualitative assessments</i>	4.61	8.82	15.34	18.73	24.89	18.70	8.92
<i>Quantitative assessments</i>	18.49	10.96	9.49	14.78	11.35	14.21	20.71
	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>
	<i>Forecasts</i>						
<i>Qualitative assessments</i>	6.73	13.81	24.84	21.48	16.97	10.19	5.99
<i>Quantitative assessments</i>	29.24	14.78	11.45	16.35	8.30	8.25	11.63
	<i>p=0.00</i>	<i>p=0.22</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>	<i>p=0.00</i>

§J. Additional rescaling of probability assessments

In Section 3 of the main text, we develop a method for scoring the accuracy of qualitative and quantitative probability assessments in equivalent terms. First, we assign each quantitative estimate to a segment of the number line corresponding to each “word of estimative probability.” Then, for every question in the data set, we calculated the mean of numeric estimates falling within each of these segments. For the purposes of this analysis we call those means *interpolated probabilities*. We then replace all numeric estimates with their corresponding interpolated probabilities as well. Otherwise, numeric estimates would demonstrate greater possible variance, and we would not be scoring qualitative/quantitative assessments on a level playing field.

In the main text of the paper, we present results of scoring interpolated probabilities using Brier Scores. But this is only one method of scoring probability estimates. In Table S8, we show how our results are robust to the logarithmic scoring rule, and also to different methods of interpolation. The columns of Table S8 represent the results of scoring estimates according to:

- Interpolated probabilities, Brier scoring
- Interpolated probabilities, Logarithmic scoring
- Words/numbers rounded to the midpoint of each segment, Brier scoring
- Qualitative assessments interpreted according to “words of estimative probability” definitions from Mosteller and Youtz (1990); quantitative estimates rounded to the nearest such definition

- Probabilities interpolated according to the segments on the 2015 Director of National Intelligence definitions of “words of estimative probability” (see Figure 1 in the main text) instead of the lexicon we used based on National Intelligence Estimates⁶²

For each approach, and for both our National Security Officials and AMT samples, we present mean respondent scores for quantitative assessments, mean respondent scores for qualitative assessments, the proportional difference between these means, and the statistical significance of this difference according to a two-way t-test. Thus we restrict our sample size to one observation per respondent, based on respondents’ mean scores rather than scoring each assessment as an independent observation. Note that lower scores indicate better assessments under Brier scoring, whereas higher scores indicate better assessments under logarithmic scoring.

This analysis shows that, regardless of interpolation method, scoring rule, and sample, quantitative assessments are less accurate than qualitative assessments throughout our data.

⁶² One drawback of the DNI’s recommendation is that it includes round numbers (e.g., 5 percent, 20 percent) as the endpoints of adjacent bins. For the purposes of this analysis, we round these numerical estimates towards 50 percent in all cases – this reduces respondent certitude and thus reduces the chances of confirming our finding that quantitative assessment exacerbates overconfidence.

Table S8. Evaluating qualitative/quantitative assessments using different operationalizations

	Interp. Probabilities, Brier Scoring (main result)	Interp. Probabilities, Logarithmic Scoring	Probs. Rounded to WEP midpoints, Brier Scoring	Probs. Rounded to nearest WEP definition via Mosteller/Youtz	Interp. Probs. Using 2015 DNI Guidelines, Brier Scoring
<i>National Security Officials (208 respondents)</i>					
<i>Qualitative assessments</i>	0.230 (.03)	-0.677 (.10)	0.231 (.03)	0.243 (.04)	0.243 (.04)
<i>Quantitative assessments</i>	0.265 (.04)	-0.812 (.14)	0.261 (.04)	0.259 (.04)	0.264 (.04)
<i>Proportion diff.</i>	13%	17%	11%	6%	8%
	$p<0.001$	$p<0.001$	$p<0.001$	$p=0.006$	$p<0.001$
<i>AMT Respondents (1,561 respondents)</i>					
<i>Qualitative assessments</i>	0.276 (.04)	-0.790 (.11)	0.273 (.04)	0.296 (.05)	0.293 (.05)
<i>Quantitative assessments</i>	0.310 (.06)	-0.934 (.19)	0.305 (.05)	0.307 (.05)	0.309 (.06)
	11%	15%	10%	4%	5%
	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$

Table S8 presents mean scores (with standard deviations) across different treatment conditions, evaluated according to different scoring rules. Statistical significance reflects two-way t-test for difference in means.

§K. Interactions between quantitative assessments and respondent attributes

Tables S9a and S9b revisit predictors of respondent Brier Scores by adding interaction terms to the models presented in the main text. The goal of this analysis is to understand whether the impact of quantifying probability assessments is consistently stronger or weaker within different subgroups of respondents.

The variables considered in this analysis are numeracy, military service, gender, U.S. citizenship, and English as a first language, as well as education level for AMT respondents. None of these interaction terms is statistically significant.

Table S9a. Examining Interaction Terms Predicting National Security Officials' Brier Scores

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
<i>Quantitative Assessment</i>	0.032 (.01) ^{***}	0.045 (.01) ^{***}	0.041 (.01) ^{***}	0.034 (.01) ^{***}	0.016 (.02)	0.017 (.01)
<i>Certitude</i>	0.032 (.02)	0.032 (.02)	0.033 (.02)	0.032 (.02)	0.033 (.02)	0.032 (.02)
<i>Forecast</i>	0.223 (.02) ^{***}					
<i>Numeracy</i>	-0.003 (2.5e ⁻)	0.000 (3.1e ⁻³)	-0.003 (2.4e ⁻)	-0.003 (2.5e ⁻)	-0.003 (2.5e ⁻)	-0.003 (2.5e ⁻)
<i>Military Service</i>	-0.001 (.01)	-0.001 (.01)	0.004 (.01)	-0.001 (.01)	-0.001 (.01)	-0.001 (.01)
<i>Female</i>	0.003 (.01)	0.002 (.01)	0.003 (.01)	0.012 (.01)	0.002 (.01)	0.002 (.01)
<i>U.S. Citizen</i>	0.011 (.01)	0.011 (.01)	0.010 (.01)	0.011 (.01)	0.004 (.01)	0.011 (.01)
<i>English Native Lang.</i>	-0.020 (.01)	-0.018 (.01)	-0.018 (.01)	-0.019 (.01)	-0.022 (.01)	-0.029 (.01)
<i>Numeracy x Quant.</i>		-0.005 (4.7e ⁻)				
<i>Service x Quant. Assessment</i>			-0.012 (.01)			
<i>Female x Quant. Assessment</i>				-0.016 (.01)		
<i>U.S. Citizen x Quant.</i>					0.019 (.02)	
<i>English x Quant.</i>						0.017 (.02)
<i>Constant</i>	0.097 (.02) ^{***}	0.090 (.02) ^{***}	0.093 (.02) ^{***}	0.096 (.02) ^{***}	0.106 (.02) ^{***}	0.107 (.02) ^{***}
<i>R²</i>	0.242	0.242	0.242	0.242	0.242	0.242

Ordinary least squares regression predicting Brier Scores, with question-level fixed effects (not shown) and standard errors clustered by respondent. All models have 7,280 observations. Note that lower Brier Scores indicate more accurate assessments. *p<.05, **p<.01, ***p<.001

Table S9b. Examining Interaction Terms Predicting AMT Respondents' Brier Scores

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
<i>Quantitative Assessment</i>	0.020 (4.4e ⁻)	0.014 (2.1e ⁻)	0.017 (3.2e ⁻)	-0.006 (.02)	0.003 (.01)	0.017 (.01)**
<i>Certitude</i>	0.297 (.01)***	0.297 (.01)***	0.297 (.01)***	0.297 (.01)***	0.297 (.01)***	0.297 (.01)***
<i>Forecast</i>	-0.237	-0.237	-0.237	-0.237	-0.237	-0.237
<i>Numeracy</i>	-0.002 (1.0e ⁻)	-0.004 (9.7e ⁻)				
<i>Military Service</i>	0.015 (.02)	0.036 (.02)	0.016 (.02)	0.015 (.02)	0.015 (.02)	0.016 (.02)
<i>Female</i>	0.009 (2.2e ⁻)	0.009 (2.2e ⁻)	0.013 (2.5e ⁻)	0.009 (2.2e ⁻)	0.009 (2.2e ⁻)	0.009 (2.2e ⁻)
<i>U.S. Citizen</i>	0.003 (.01)	0.003 (.01)	0.004 (.01)	-0.007 (.01)	0.002 (.01)	0.002 (.01)
<i>English Native Lang. Education</i>	0.004 (.01)	0.003 (.01)	0.003 (.01)	0.002 (.01)	-0.002 (.01)	0.003 (.01)
	-0.011 (1.6e ⁻)	-0.010 (1.9e ⁻)				
<i>Numeracy x Quant.</i>	-0.003 (1.8e ⁻)					
<i>Service x Quant. Assessment</i>		-0.043 (.04)				
<i>Female x Quant. Assessment</i>			-0.008 (4.4e ⁻)			
<i>U.S. Citizen x Quant.</i>				0.019 (.02)		
<i>English x Quant.</i>					0.010 (.02)	
<i>Education x Quant.</i>						-0.002 (3.2e ⁻)
<i>Constant</i>	0.214 (.01)***	0.218 (.01)***	0.215 (.01)***	0.229 (.01)***	0.224 (.01)***	-0.216
<i>R²</i>	0.221	0.221	0.221	0.221	0.221	0.221

Ordinary least squares regression predicting Brier Scores, with question-level fixed effects (not shown) and standard errors clustered by respondent. All models have 54,621 observations. Note that lower Brier Scores indicate more accurate assessments. *p<.05, **p<.01, ***p<.001

§L. Replication of Survey Experiment 2 results using logarithmic scoring

The main text of the paper analyzes variation in probability assessment scores across respondents.

Table S10 replicates that analysis, evaluating respondent assessments using logarithmic scoring instead of Brier scoring. Note that higher logarithmic scores indicate better assessment.

Table S10. Relationship between quantitative assessment, respondent certitude, and accuracy (logarithmic scoring)

National Security Officials

	Model 1	Model 2
<i>Quantitative Assessment</i>	-0.104 (.02)***	0.022 (.02)
<i>Certitude</i>	-0.530 (.07)***	-0.237 (.10)*
<i>Quantitative Assessment x</i>		-0.503 (.12)***
<i>Forecast</i>	-0.611 (.06)***	-0.606 (.06)***
<i>Numeracy</i>	0.010 (.01)	0.010 (.01)
<i>Military Service</i>	0.006 (.02)	0.006 (.02)
<i>Female</i>	0.001 (.02)	-0.000 (.02)
<i>U.S. Citizen</i>	-0.029 (.04)	-0.041 (.04)
<i>English Native Lang.</i>	0.061 (.04)	0.072 (.04)*
<i>Education Level</i>		
<i>Constant</i>	-0.208 (.05)***	-0.271 (.05)***
<i>R²</i>	0.210	0.212

Ordinary least squares regression predicting logarithmic scores for individual probability assessments. Both models have 7,280 observations. Robust standard errors clustered by respondent. Question-level fixed effects not shown. Note that higher logarithmic scores indicate more accurate assessments. *p<.05, **p<.01, ***p<.001.

AMT Respondents

	Model 3	Model 4
<i>Quantitative Assessment</i>	-0.051 (.01)***	0.026 (.01)***
<i>Certitude</i>	-1.317 (.03)***	-1.101 (.04)***
<i>Quantitative Assessment x</i>		-0.358 (.05)***
<i>Forecast</i>	-0.680 (.02)***	-0.682 (.02)***
<i>Numeracy</i>	0.012 (2.9e ⁻³)***	0.012 (2.9e ⁻³)***
<i>Military Service</i>	-0.046 (.06)	-0.036 (.06)
<i>Female</i>	-0.022 (.01)***	-0.020 (.01)**
<i>U.S. Citizen</i>	-0.007 (.03)	-0.004 (.03)
<i>English Native Lang.</i>	-0.010 (.03)	-0.010 (.03)
<i>Education Level</i>	0.032 (4.9e ⁻³)***	0.032 (4.9e ⁻³)***
<i>Constant</i>	-0.536 (.04)***	-0.581 (.04)***
<i>R²</i>	0.244	0.245

Ordinary least squares regression predicting logarithmic scores for individual probability assessments. Both models have 54,621 observations. Robust standard errors clustered by respondent. Question-level fixed effects not shown. Note that higher logarithmic scores indicate more accurate assessments. *p<.05, **p<.01, ***p<.001.