It Matters Whether Probabilities are Expressed in Numbers Versus Words: Experimental Evidence from National Security

Faculty Research Working Paper Series

Jeffrey A. Friedman
Dartmouth College

Jennifer S. Lerner
Harvard Kennedy School

Richard Zeckhauser
Harvard Kennedy School

April 2016

RWP16-016

It Matters Whether Probabilities are Expressed in Numbers Versus Words: Experimental Evidence from National Security Professionals

Jeffrey A. Friedman, *Assistant Professor of Government, Dartmouth College*[1]
Jennifer S. Lerner, *Professor of Public Policy and Management, Harvard University*
Richard Zeckhauser, *Frank P. Ramsey Professor of Political Economy, Harvard University*

Working Paper: December 15, 2015
9,932 words

Abstract. National security is one of many fields where public officials offer imprecise probability assessments when evaluating high-stakes decisions. This practice is often justified with arguments about how quantifying subjective judgments would bias analysts and decision makers toward overconfident action. We translate these arguments into testable hypotheses, and evaluate their validity through survey experiments involving national security professionals. Results reveal that when decision makers receive numerals (as opposed to words) for probability assessments, they are less likely to support risky actions and more receptive to gathering additional information, disconfirming the idea of a bias toward action. Yet when respondents generate probabilities themselves, using numbers (as opposed to words) magnifies overconfidence, especially among low-performing assessors. These results hone directions for research among both proponents and skeptics of quantifying probability estimates in national security and other fields. Given that uncertainty surrounds virtually all intelligence reports, military plans, and national security decisions, understanding how national security officials form and interpret probability assessments has wide-ranging implications for theory and practice.

Introduction

When General Stanley McChrystal wrote to President Barack Obama in 2009, recommending that 40,000 additional soldiers be deployed to Afghanistan, he explained that his proposal offered "acceptable risk," presenting "the best prospect for success in this important mission." Yet McChrystal's report did not specify what an "acceptable risk" entailed. Moreover, even if some action offers the best chances of success, this does not imply that those chances are worth accepting. Though readers could not logically evaluate the Afghan Surge without estimating how much it increased the U.S. military's chances of success, McChystal's report left this judgment unclear.

Similarly vague probability assessments appear regularly, and deliberately, in national security decision making. One study of declassified National Intelligence Estimates written from 1964-1994 found that 96 percent of key judgments expressed probability in ways that lacked clear quantitative equivalents (Friedman and Zeckhauser 2012). Figures 1 and 2 present official guidelines for assessing uncertainty in intelligence, describing risk in military planning, and evaluating the reliability of human sources. These guidelines instruct analysts to express probability using qualitative terms.

[1] jeffrey.a.friedman@dartmouth.edu. Joowon Kim, Ben Rutan, and Max Yakubovich provided exceptional research assistance. For comments on drafts and research design, we thank Paul Novosad, Brendan Nyhan, Christopher Robert, Bryan Pendleton, Kathryn Schwartz, and Peter Scoblic. A previous version of this paper was presented to Dartmouth's Government Department. This research was funded by the Department of Homeland Security via the University of Southern California's Center for Risk and Economic Analysis of Terrorism Events (CREATE). All errors are the authors'.

[Figures 1 and 2]

National security officials are not unique in their skepticism towards expressing probability assessments precisely. Regulation (Sunstein 2007), environmental policy (Budescu, Broomell, and Por 2014), and medicine (Gigerenzer 2002) represent other areas of high-stakes decision making where quantifying probability estimates engenders controversy. Across these fields, many scholars and practitioners argue that quantifying probability estimates would produce harmful behavioral consequences.

This paper focuses on two common arguments to this effect, which we label *Illusions of Rigor* and *Numbers as a Second Language*. In brief, the "Illusions of Rigor" argument holds that if national security analysts quantified probability assessments, then decision makers would interpret those assessments as being more rigorous than they really are. The "Numbers as a Second Language" argument holds that asking analysts to translate subjective beliefs into numeric estimates would be unnatural, and could thus induce avoidable errors.

Both arguments imply that quantifying subjective probabilities, however justifiable in principle, would have meaningful drawbacks in practice. Scholars have debated these claims for more than fifty years (Kent 1964). Given widespread cognitive constraints on national security decision making (McDermott 1998, Johnson 2004, Yarhi-Milo 2014, Rapport 2015), the Illusions of Rigor and Numbers as a Second Language arguments are certainly plausible. Yet that does not mean that these arguments should also be accepted at face value, and we are not aware of any research testing these claims directly. We therefore hone the "Illusions of Rigor" and "Numbers as a Second Language" arguments into testable hypotheses which we evaluate through two original survey experiments administered to 407 national security officials and 3,017 respondents from Amazon Mechanical Turk.

Our results disconfirm the "Illusions of Rigor" argument. Contrary to conventional wisdom, we find that quantifying probability assessments makes decision makers less willing to support risky action, and more receptive to gathering additional information. Yet our results also support a particular version of the "Numbers as a Second Language" argument: when respondents estimated probabilities themselves, quantification magnified overconfidence, especially among low-performing assessors. These results hardly close debates about the costs and benefits of vague probability assessments, but they do provide new empirical evidence about the behavioral consequences of quantifying subjective probabilities.

The paper proceeds in five sections. Section 1 reviews the evidence-based benefits of quantifying probability assessments in foreign policy analysis and other areas of high-stakes decision making. Section 2 reviews the "Illusions of Rigor" and "Numbers as a Second Language" arguments, explaining how these arguments lack theoretical specification and empirical support. Section 3 presents our first survey experiment, examining how quantification influences the way decision makers respond to probability assessments. Section 4 presents our second survey experiment, examining how quantification influences the way analysts estimate probabilities. Section 5 concludes.

Section 1. The Case For Quantifying Probability Assessments

Probability assessments describe the likelihood that some outcome will occur or that some statement is true.[2] Probability assessments apply both to predictions and to unknown current or prior events. Virtually all national security decisions depend on probability assessments, even if national security officials express those judgments vaguely. For this reason, assessments of uncertainty play a central role in international relations theory (Rathbun 2007), formal models of coercive bargaining (Powell 2002), and professional military literature (Connable 2012).

There are four principal, evidence-based benefits to quantifying probability assessments, even when those assessments rely on subjective judgment rather than mathematical or statistical reasoning. The first of these benefits is avoiding miscommunication. Terms such as "a fair chance" or "acceptable risk" mean different things to different people (Beyth-Merom 1982, Mosteller and Youtz 1990). Even when readers receive guidelines defining such terms unambiguously, they often still interpret qualitative probability assessments in ways that authors did not intend (Budescu et al. 2014). Individuals have natural inclinations to interpret vague probability assessments in self-serving fashions (Piercey 2009) and political incentives to exploit ambiguity in order to justify controversial decisions (Rovner 2011).

Second, quantifying probability assessments provides information that qualitative expressions obscure. The U.S. Intelligence Community's current guidelines for expressing probability cannot distinguish assessments of 1-in-5 versus 1-in-20 (or 1-in-20 versus 1-in-100), even though these distinctions can substantially influence high-stakes decision making. Moreover, research shows that foreign policy analysts can meaningfully discriminate among probabilities within these ranges (Friedman et al. 2015). Quantification reveals discrepancies among analysts' views (Kent 1964), encouraging productive debate about why estimates differ and whether further collection and analysis is warranted to resolve those disagreements.

Third, quantifying probability assessments affords greater flexibility to update beliefs in light of new information. Mellers et al. (2014) show that geopolitical forecasters are most effective when they update their beliefs frequently and in small increments, for example shifting an estimate of 10 percent to 12 percent. "Words of estimative probability" inhibit making and communicating these adjustments.

Fourth, quantifying probability assessments promotes accountability, evaluation, and improvement (Tetlock and Mellers 2011, Dhami et al. 2015). National security analysts are often accused of using "weasel words" to escape criticism (Lowenthal 1999, 87). Even subconsciously, many people reinterpret their probability assessments after the fact in a manner that supports illusions of good judgment (Dawes 1988, Tetlock 1999). By contrast, quantitative probability assessments facilitate objective evaluations of, and often substantial improvements

[2] "Probability" is distinct from "confidence." The former conveys an estimate of likelihood; the latter describes the strength of that judgment. For example, the probability that a fair coin flip comes up heads is 50 percent and an analyst should be highly confident in this estimate. The distinction between likelihood and confidence is encoded in official intelligence guidance, yet the terms are often conflated in practice (Friedman and Zeckhauser 2015). One common objection to quantifying probability assessment is that this implies undue confidence, but this is a reason to assess likelihood and confidence separately, not to merge distinct concepts. (Note that the term "overconfidence" does have an application to probability assessment, which we discuss below.)

to, analysts' performance (Alpert and Raiffa 1982, Rieber 2004, Mellers et al. 2014, Dhami et al. 2015).

Vague probability assessments sacrifice these benefits. Preferring the qualitative expression of uncertainty thus requires explaining why probabilistic precision incurs costs that outweigh the arguments reviewed in this section.

Section 2. The Case Against Quantifying Probability Assessments

One common objection to quantifying probability assessments is the notion that it is impossible to coherently translate subjective beliefs into numeric expressions. This objection is unfounded. Any probabilistic belief, no matter how subjective, can be coherently communicated with a single numeric percentage (Savage 1954). Moreover, research programs such as the Good Judgment Project have shown that foreign policy analysts can make numeric probability assessments effectively (Mellers et al. 2014, Tetlock and Gardner 2015).

Credible objections to probabilistic precision must therefore invoke assumptions about how quantifying subjective assessments generates nonrational biases that harm the quality of decision making or analysis. We define "decision makers" as individuals who *interpret* probability assessments and "analysts" as individuals who *provide* probability assessments. In this section, we describe how quantifying probability estimates could potentially influence each of these groups. We also explain how commonly-voiced arguments in this area currently lack both theoretical specification and empirical backing, especially in the national security context.

"Illusions of Rigor" when interpreting probabilities

Many national security scholars and practitioners warn that expressing probability assessments precisely would make subjective judgments appear more rigorous than they really are (Weiss 2008, Marchio 2014). The resulting "Illusions of Rigor" could lead national security decision makers to neglect their analytic limitations.

Though this concern is commonly stated, it is rarely translated into testable hypotheses. For example, if General McChrystal had told President Obama that there was a "35 percent chance" of the Afghan Surge succeeding, would this have made the decision seem more defensible than framing it as an "acceptable risk"? Or would the extra precision have highlighted the prospect of failure and thus made the President *less* likely to back McChrystal's plan?

To our knowledge, no national security scholars have translated the "Illusions of Rigor" argument into falsifiable hypotheses, let alone confirmed those hypotheses empirically. Thus Section 3 lays out four specific ways that quantifying subjective probabilities could influence the way decision makers evaluate risky actions. Each of these mechanisms could plausibly reduce the quality of national security decision making. Yet without testing these hypotheses directly, objections to the implied rigor of numeric estimates remain speculative.

"Numbers as a Second Language" when estimating probabilities

Even if national security decision makers respond rationally to numeric probabilities, quantifying subjective judgments could still be detrimental if this alters the content of the information that analysts provide. Some scholars argue that analysts naturally think about uncertainty qualitatively (Zimmer 1994, Wallsten and Budescu 1995). Quantitative expressions of probability thus resemble speaking in a second language, a task that could induce avoidable errors.

This "Numbers as a Second Language" argument also lacks rigorous substantiation, especially in a national security context. Skeptics might even argue that this proposition is untestable, as there is often no objective way to compare the accuracy of qualitative and quantitative assessments. Interpretations of qualitative probability estimates vary across subjects and contexts (Beyth-Merom 1982, Mosteller and Youtz 1990). Generally speaking, scholars cannot be sure exactly how an analyst who provided a "likely" estimate should have quantified that assessment, or how an analyst who provided a "65 percent" estimate should have described that judgment qualitatively.

The "words of estimative probability" guidelines shown in Figure 1 provide rare traction here, because they represent an effort by the U.S. Intelligence Community to establish context-neutral mappings between qualitative and quantitative probability assessments. The assessments "likely" and "65 percent," for example, are clearly equivalent according to these guidelines. This structures comparisons of qualitative and quantitative assessments, but also points to a potential weakness of the "Numbers as a Second Language" argument. Once national security officials define "words of estimative probability" in relation to the number line, why should we expect qualitative and quantitative assessments to differ meaningfully? Section 4 describes and tests four hypotheses on this issue.

Empirical approach

Evaluating the behavioral consequences of quantifying probability assessments thus requires addressing two distinct, empirical questions. First, do decision makers evaluate risky actions differently when those actions are described with quantitative versus qualitative probabilities? Second, do analysts provide less accurate probability assessments when they quantify those judgments? The remainder of this paper describes two survey experiments designed to address these questions.

We administered surveys to a total of 407 national security officials enrolled in two advanced military education programs. These officials were predominantly active-duty U.S. military officers at the ranks of Lieutenant Colonel and Colonel (for the U.S. Army, Air Force, and Marines), or the equivalent U.S. Navy ranks of Commander and Captain. In the U.S. military, advanced education programs are mandatory for promotion to the rank of Colonel/Captain. Thirteen percent of these officials were similarly-ranked military officers from other countries. Twenty-five percent were civilians from the U.S. Intelligence Community, Department of State, and other agencies. Supplementary material describes respondent demographics in more detail.

To bolster the internal validity of our studies, we also administered surveys to a total of 3,017 respondents via Amazon Mechanical Turk (MTurk).[3] As shown below, elite (national security) and general (MTurk) samples of respondents generate complementary results. This supports the robustness and generalizability of our findings, while also building plausibility for applying broader decision science research to the specific problems facing national security officials (Hyde 2015, Renshon 2015). Following experimental treatments, we asked respondents to complete a three-question adaptive Berlin Numeracy Test (Cokely et al. 2012) and a standard battery of demographic questions.

Section 3. How Decision Makers Interpret Probability Assessments

If quantifying probability assessments creates inappropriate "Illusions of Rigor," this should influence decision making in at least one of the four following ways.

Hypothesis 1: Quantifying probability assessments makes decision makers more likely to support risky actions. Quantifying subjective judgments could alleviate decision makers' concerns about placing personnel, resources, and national interests at risk based on incomplete information. If "Illusions of Rigor" cause decision makers to see quantitative assessments as being sounder than they really are, this could make risk-taking seem more defensible, thus encouraging decision makers to support risky actions.

Hypothesis 2: Quantifying probability assessments amplifies decision makers' evaluations of risky actions. Quantifying favorable probability assessments could cause decision makers to believe that they have an unusually sound basis for taking risks. Yet quantifying *un*favorable probability assessments could make some actions appear excessively risky. It is thus possible that quantifying subjective probabilities amplifies decision makers' tendencies to support or oppose a proposed policy.

Hypothesis 3: Quantifying probability assessments reduces decision makers' willingness to gather additional information when evaluating risky actions. This is the most straightforward implication of the idea that numeric probabilities seem more rigorous than they really are. If numeric probabilities cause national security officials to overestimate the soundness of the analysis underlying their decisions, they might rush to judgment rather than gathering additional information.

Hypothesis 4: Quantifying probability assessments raises decision makers' confidence levels when evaluating risky actions. Even if quantifying subjective probability does not influence levels of support for risky action or willingness to gather additional information, "illusions of rigor" could cause decision makers to hold undue faith in their abilities to evaluate uncertain choices. This could harm the quality of decision making by leading decision makers to over- or under-resource the policies they select.

[3] We conducted surveys with both National Security Officials and MTurk samples between August 5-7, 2015. MTurk respondents were required to be U.S. residents at least 18 years of age. They were compensated $2.04 for completing the survey.

Survey design

We tested Hypotheses 1-4 by presenting respondents with fictional vignettes involving national security decisions under uncertainty. Appendix A provides the text of these vignettes, comprising a hostage rescue mission in Syria, a drone strike in Yemen, and supporting local security forces in Afghanistan. Scenarios were selected to be plausible to military officers and familiar to a broad audience. After each vignette, we asked respondents how strongly they supported the proposed action, how strongly they supported waiting for additional information before deciding, and how confident they were in making their assessments. We elicited these evaluations on seven-point scales.[4]

The use of fictional vignettes to examine prospective reactions to real-life decisions has clear limitations, but follows recent international relations scholarship such as Press, Sagan, and Valentino (2013), Tomz and Weeks (2013), and Kertzer and Brutger (forthcoming). Indeed, when gauging support for prospective national security policies, scholars commonly examine how respondents react to probability assessments, such as estimates of the chances that an action will succeed. Understanding whether different methods of communicating probability influence responses to that information thus carries implications for designing experimental studies, as well as for informing practical debates about the conduct of national security analysis and decision making.

We randomly assigned respondents to qualitative and quantitative assessment conditions. In the qualitative assessment condition, all probability assessments within scenarios were expressed using one of the seven "words of estimative probability" shown at the top of Figure 1.[5] In the quantitative assessment condition, we converted those words into percentages using the probability closest to the middle of the range that each "word of estimative probability" represented, rounded to multiples of 0.05. Thus, we converted "even chance" to "50 percent," we translated "likely" to "65 percent," and so forth.

We administered this survey to 208 participants in an advanced military education program. For the remainder of the paper we refer to this group as our "National Security Officials Sample." We also administered this survey to 1,458 respondents on Amazon Mechanical Turk. In both groups, we randomized the order in which the three scenarios appeared. Respondents viewed one scenario at a time, and could not change responses after moving to the next vignette.

We also randomly presented respondents with one of three versions of each scenario. We varied probability assessments across vignette versions to represent what we considered to be "optimistic," "neutral," or "pessimistic" information about proposed actions. For example, in the "optimistic" version of the hostage rescue vignette, intelligence analysts estimated that there was an 80 percent chance ("very likely") that the hostages were at the suspected location. This assessment was placed at 65 percent ("likely") and 50 percent ("even chance") for the neutral and pessimistic versions of this scenario, respectively. Supplementary material shows how probability assessments varied across vignette versions.

[4] We also asked respondents to "write a few sentences" justifying their views.
[5] We selected this lexicon over the more recent version released by the U.S. Director of National Intelligence for three main reasons: the original spectrum appears more qualitative in nature; it is easier to embed in online survey platforms; and dividing the number line into seven equal segments places integer percentages into unique bins.

Varying probability assessments across vignettes enabled us to confirm that these probability assessments influenced how respondents evaluated the actions we proposed. Table 1 shows mean levels of support for proposed actions across different vignette versions. In all cases, respondents demonstrated consistently greater support for risky actions when those actions were described with more favorable probability assessments. On the whole, respondents generally opposed risky actions described with what we considered to be "Pessimistic" assessments, and they generally supported actions described with what we considered to be "Optimistic" assessments.

[Table 1]

To add robustness to our elite sample results, we administered a shorter survey, containing only the "neutral" hostage vignette, to 199 students in another advanced military education program. We refer to this supplementary survey experiment as "Elite Sample B." Altogether, we administered Experiment 1 to 1,857 respondents, who evaluated a total of 5,173 scenarios.

Testing Hypothesis 1

Hypothesis 1 predicts that quantifying probability assessments should make decision makers more likely to support risky actions. Tables 2a and 2b below examine this hypotheses. The dependent variable is respondent support for proposed actions (7-point scale). *Quantitative Assessment* is a dummy variable, taking the value 1 if respondents were assigned to read scenarios involving numeric probabilities, and 0 if not. Control variables include indicators for whether vignettes were *Optimistic* or *Pessimistic*; indicators for the vignette's subject (*Hostage Scenario*, *Drone Scenario*); respondent *Numeracy* (4-point scale); and indicators for whether respondents were female, U.S. citizens, military officers, and native English speakers.[6] We model support for risky action using ordinary least squares regression with respondent fixed effects.[7]

[Tables 2a and 2b]

Our results show that quantifying probability estimates does not make respondents more likely to support risky action. In both cases, respondents were, in fact, less likely to support actions described with numeric probabilities. This effect is statistically significant ($p<0.001$) in the MTurk data and Elite Sample B.[8] Hypothesis 1 thus fails three experimental tests including two elite samples and a larger pool of MTurk respondents. In supplementary material, we analyze these patterns further, examining results within all nine individual vignettes. We found no instance where quantifying probability assessments consistently increased support for risky action.

[6] In the MTurk sample, the *Military Service* variable indicates respondents with any current or previous military service. Less than 1% of MTurk respondents were active-duty military personnel.
[7] In supplementary material, we show that our results hold using ordered probit analysis.
[8] 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.

Testing Hypothesis 2

Hypothesis 2 predicts that quantifying probability assessments amplifies policy evaluations. If this is true, we should observe two findings. First, an interaction term between quantitative assessment and a dummy for our "pessimistic" vignettes should have a negative coefficient. This would show that quantifying probabilities makes bad options seem worse. The second columns of Tables 2a and 2b show that the data do not sustain this prediction. In both cases, the coefficient on this interaction term is positive.

Next, Hypothesis 2 predicts that an interaction term between quantitative assessment and a dummy for our "optimistic" vignettes should have a positive coefficient that outweighs the degree to which quantifying probability assessments reduces support for risky action overall. This would indicate that quantifying probability assessments makes good choices seem better. The data do not support this prediction. In both samples, the interaction term between quantitative assessments and "optimistic" scenario versions does not outweigh the estimated direct effect of quantifying probabilities. Thus, even when respondents evaluated proposals described with "optimistic" assessments, quantifying probability assessments reduced support for risky action.[9]

Testing Hypotheses 3 and 4

Hypothesis 3 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. The third columns in Tables 2a and 2b show that respondents are more willing to gather additional information when presented with quantitative probability assessments. Among the 199 respondents in Elite Sample B, respondents were even more supportive of gathering additional information when presented with numeric probabilities. The mean support for delaying decision across respondents in the qualitative assessment condition was 3.14 (standard deviation 1.97). Support for delaying decision was nearly a full point higher in the quantitative assessment condition (average 4.11, standard deviation 2.07, p=0.001).

Our results also refute Hypothesis 4, which predicts that quantifying probability assessments should make decision makers more confident in their policy evaluations. In all three samples, quantifying probability assessments had no consistent impact on respondents' confidence levels.

Discussion of results

The results from Survey Experiment 1 do not suggest that quantitative and qualitative probability assessments produce identical responses from decision makers. Respondents given numeric probabilities were more cautious in supporting risky actions. This is not necessarily ideal: sometimes, the right move is to act, incomplete information and all.

Nevertheless, a large body of scholarship supports a general consensus that national security decision makers are prone to overconfidence (Johnson 2004), that they neglect to address key

[9] Again, supplementary material shows how there was no scenario version – even optimistic scenarios – where quantifying probability assessments significantly increased support for risky action.

uncertainties (Rapport 2015), and that they are overly inclined to pursue risky actions (Kahneman and Renshon 2007). One argument in favor of quantifying probability assessments in any field of high-stakes decision making is that this prevents decision makers from glossing over key uncertainties, or interpreting ambiguous information in ways that support overly risky actions. The Illusions of Rigor argument is important because it suggests that attempts to highlight uncertainty by quantifying probability assessments can backfire, unintentionally increasing decision makers' willingness to take risks on the basis of incomplete information. The data presented here do not support that argument, however, disconfirming four important hypotheses about the behavioral consequences of quantifying probability assessments.

Section 4. How Analysts Estimate Probabilities

The "Numbers as a Second Language" argument rests primarily on the claim that asking analysts to quantify their subjective beliefs adds error to the assessments they provide.

Hypothesis 5: National security analysts make less accurate probability assessments using numbers versus words. If Hypothesis 5 is confirmed, it becomes crucial to understand why quantification degrades the accuracy of probabilistic judgment. "Words of estimative probability" lexicons, such as the guidelines shown in Figure 1, map qualitative and quantitative probability assessments together. In principle, the right way to use these lexicons is to estimate a probability precisely enough to determine which segment of the number line it falls into. Yet there are at least three reasons why respondents might behave differently.

Hypothesis 6: Quantifying probability estimates reduces accuracy by adding random noise to assessments. Respondents using "words of estimative probability" may not actually specify their judgments with respect to the number line. In this sense, "words of estimative probability" might be more appropriately construed as a Likert scale, closer to a "feeling thermometer" for measuring uncertainty than a summary of some underlying, quantitative judgment. A large literature demonstrates how such response scaling can generate unpredictable behavior (Krosnick and Frabrigar 1997). From this perspective, divergences between the use of qualitative and quantitative probabilities may effectively appear as random noise.

Hypothesis 7: Quantifying probability estimates reduces accuracy by inducing underconfidence. Respondents asked to quantify subjective probabilities may be more anxious about "going out on a limb" with their estimates. For example, estimates of "65 percent" and "likely" are literally equivalent according to the "words of estimative probability" spectrums in Figure 1. However, if these estimates are applied to statements that ultimately prove false, the word "likely" can be interpreted after the fact as meaning something far less than 65 percent. Fearing such criticism, analysts may tilt to caution, and thus be excessively underconfident, when quantifying probability assessments.[10]

[10] Note that the terms "underconfidence" and "overconfidence," as used here, assess the quality of probability assessments. Assessments are underconfident when estimated probabilities are systematically less extreme than observed frequencies. Assessments are overconfident when the reverse is true. These uses of the terms "overconfidence" and "underconfidence" are distinct from evaluating whether probability assessments are "overprecise" or "underprecise." Moore and Healy 2008 parse definitions of "confidence" in more detail.

Hypothesis 8: Quantifying probability estimates reduces accuracy by inducing overconfidence. "Words of estimative probability" spectrums could also prevent excessive overconfidence by imposing natural anchors for relaying judgments. Imagine that an analyst is quite sure that a statement is true, while acknowledging residual uncertainty. Using numbers, this analyst might express her judgment as 95 percent: a high probability that also clearly leaves room for doubt. Using the guidelines shown at the top of Figure 1, our analyst might instead have chosen the term "very likely." Even though the range of probabilities this term comprises stops substantially short of 95 percent, the term "almost certain" does not clearly signal residual uncertainty. In this sense, "words of estimative probability" spectrums impose natural cut-points on judgments that may systematically reduce the certitude that analysts attach to their estimates. Especially given how most probability assessors naturally tend towards overconfidence (Alpert and Raiffa 1982, Tetlock 2005), this ratcheting effect could consistently improve assessments of uncertainty.

These hypotheses, beyond their theoretical implications for understanding the psychology of probability assessment, have important practical connotations. If the "Numbers as a Second Language" problem manifests itself as random noise, then rectifying this problem could be very difficult. Yet if quantifying estimative probabilities generates specific, predictable errors, then those errors may be correctible. We return to this subject later in the paper. For now, the important point is that testing the "Numbers as a Second Language" argument calls both for evaluating a potential relationship between quantification and estimative accuracy, and for exploring the mechanisms by which any such treatment effect is transmitted.

Survey design

We tested Hypotheses 5-8 by asking respondents to make probability estimates in response to 40 randomly-ordered questions about foreign policy and national security. The full question list appears in supplementary material.

We varied questions across three analytic types. 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?") Five questions involved 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?") We varied question types in order to examine whether systematic differences between qualitative and quantitative assessments differed across question types. As shown in supplementary material, we found no evidence of this.

We randomly assigned respondents to estimate probabilities using either numeric percentages or "words of estimative probability." We administered this survey to our National Security

Officials sample and to 1,561 respondents on Amazon Mechanical Turk.[11] These surveys produced 53,070 probability estimates that we can evaluate as of this writing.

Scoring probability estimates

We scored qualitative and quantitative estimates in equivalent terms using the following procedure. First, we calculated the mean numeric assessment corresponding to each "word of estimative probability" for each question we posed.[12] 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 estimates on a level playing field.

After translating probability estimates into equivalent terms, we evaluated their accuracy using Brier Scores.[13] Using this method, we found that 74 percent of MTurk respondents and 95 percent of national security officials provided assessments that were more informative, on average, than random guessing.[14] This indicates that a large majority of participants took the probability estimation exercise seriously, especially given that even experts often struggle to beat the "as-good-as-random" standard when evaluated with proper scoring rules (Tetlock 2005).

Evaluating Hypothesis 5

Figure 3 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 13 percent. For MTurk respondents the equivalent gap was 11 percent. Both comparisons were highly statistically significant (p<0.001).

[Figure 3]

Alternative scoring rules produced similar patterns. With logarithmic scoring, the difference across treatment conditions in respondents' mean Brier Scores was 19 percent for National

[11] 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. MTurk respondents were randomly assigned to complete only one of our two experiments.

[12] The "words of estimative probability" spectrum we used divides the number line into seven equal segments, and thus each integer percentage falls within a unique bin. Though the 2015 WEP spectrum presented by the DNI defines bins more explicitly, it also leaves ambiguous as to how estimates such as "5 percent" or "80 percent" should be encoded.

[13] Brier Scores represent 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 proves false, then the respondent's Brier Score for that prediction is $(0-p)^2$. Lower Brier Scores represent more accurate assessments.

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

Security Officials and 18 percent for MTurk respondents.[15] 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 11 percent for National Security Officials and 10 percent for MTurk respondents.[16]

Table 3 shows how a dummy variable indicating assignment to the *Quantitative Assessment* treatment condition correlates strongly with respondents' mean Brier Scores in multivariate analysis. We control for respondent's *Numeracy* score (4-point scale), along with dummy variables indicating whether respondents were female, whether they were currently (or formerly) members of their country's military, whether they were U.S. citizens, and whether English was their native language. For MTurk respondents, we also include variables for respondents' education level.[17]

[Table 3]

One critique 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 14 percent. Yet 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.[18] First, we scored qualitative estimates according to Mosteller and Youtz's (1990) meta-study of how respondents typically evaluate these terms, and we rounded numeric estimates to the nearest of these anchors.[19] This method found that quantitative assessors still were less accurate, producing mean Brier scores that were 9 percent worse among National Security Officials (p=0.002) and 5 percent worse among MTurkers (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 of extreme estimates. This approach showed that the degradations in performance associated with quantitative estimation were 6% and 4% among National Security Official and MTurk respondents, respectively.[20] Thus while it is important to acknowledge the difficulty of evaluating qualitative probability assessments even when instructing respondents to make those

[15] 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.
[16] See supplement for additional analysis.
[17] Education is measured on a 3-point scale: high school degree or less, college degree (2 or 4-year), postgraduate training. A variable for respondent age did not approach significance and eliminated 95 respondents due to missing data. As in previous analysis, the *Military Service* variable captures current military officers in the National Security Officials sample; among MTurk respondents, this variable indicates any current or prior military service.
[18] See supplement for additional analysis.
[19] 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."
[20] See supplementary material for more information.

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.

Evaluating Hypotheses 6-8

Table 4 compares the frequency with which qualitative and quantitative assessors provided estimates corresponding to each "word of estimative probability." Within both elite and general samples, respondents who used numeric probabilities gave estimates comprising noticeably greater certitude. For example, respondents were more than three times as likely to make "remote" probability assessments when using numbers rather than words.

[Table 4]

The fact that quantitative assessors displayed greater certitude does not guarantee that their estimates will be worse. After all, an omniscient assessor would have complete certitude. However, our data suggest support for Hypothesis 8: that quantifying probability estimates induces overconfidence, and that this drives the gap in performance between qualitative and quantitative assessors in our study.

To evaluate this relationship statistically, we calculated each respondent's mean level of *Certitude*, defined as one minus the Brier Score that a respondent would expect to receive if her predictions were perfectly calibrated.[21] Table 5 presents ordinary least squares regression showing that Quantitative Assessment strongly predicts respondents' Certitude. We then show that controlling for respondents' Certitude reduces the partial correlation between Quantitative Assessment and respondents' mean Brier Scores.

[Table 5]

Following the causal mediation analysis technique of Imai, Keele, and Tingley (2010), we estimate that Certitude mediates an estimated 19 percent of the treatment effect between the use of numeric probabilities and the accuracy of resulting assessments among National Security Officials and 82 percent of this treatment effect among MTurk respondents.[22] As shown in supplementary material, estimates of causal mediation are higher (47 percent for National Security Officials and 89 percent for MTurk respondents) when we evaluate probability assessment using logarithmic scoring, which is more sensitive than the Brier Score to the calibration of probability estimates near zero and one.

Sensitivity analysis

We found no indication that the performance gap between qualitative and quantitative assessments was driven by respondents' numeracy, gender, language, nationality, age, education, or military experience. As shown in supplementary material, interaction terms between Quantitative Assessment and these factors are statistically insignificant.

[21] Thus we define a probability estimate's *Certitude* (C) as $C(p) = 1 - p(1-p)^2 + (1-p)(0-p)^2$. This approach accounts for the Brier Score's nonlinear risk/reward tradeoffs.
[22] See supplement for sensitivity analysis of mediation estimates.

Of the subsets of respondents who participated in our study, the one where the decrement associated with quantitative assessment is by far the largest is for the worst-performing probability assessors. 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 6 percent among National Security Officials (p=0.02) and to 3 percent among MTurk respondents (p=0.001). If we limit the analysis to respondents whose Brier Scores were better than the median within their respective samples, then mean performance for quantitative and quantitative MTurk assessors is identical to the fifth decimal place (p=1.00), while the 3 percent degradation in performance among numerical assessors in the National Security Officials sample is not statistically significant (p=0.23).

Discussion of results

The results of Survey 2 support the argument that national security analysts can more easily calibrate probabilistic judgments using qualitative versus quantitative expressions. This pattern is at least partially driven by a general tendency for quantitative assessors to attach additional (and excessive) certitude to their judgments. We found that this problem appeared mainly among the worst probability assessors in our data set. These findings hone broad concerns about quantitative assessments of uncertainty being a "second language" into a specific causal mechanism. We conclude by discussing the feasibility of correcting this problem, along with other implications of our findings and directions for future research.

Section 5. Directions for future research

This paper addresses longstanding debates about the desirability of quantifying probability assessments in high-stakes decision making, particularly in the national security domain. We distilled existing objections to this practice into eight testable hypotheses. We then evaluated those hypotheses with two original survey experiments administered to national security officials and to 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.

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 risky 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.

We expect that appropriate training can counteract this bias. For example, the Good Judgment Project has rigorously demonstrated that even one-hour training sessions in probability assessment bring marked improvement. Four years of data from this research program show that, in general, experienced probability assessors do not demonstrate the kinds of overconfidence that we observed in our experiments.[23] On the whole, Good Judgment Project analysts were slightly

[23] See the sources cited in Section 1.

*under*confident in their assessments, a finding replicated in a recent large-scale study of declassified intelligence estimates (Mandel and Barnes 2014). And because the treatment effect we observed in our study originated mainly with the worst assessors in the data, we suspect that our results overstate the extent to which this problem would appear among professional assessors who possess greater training, experience, and incentives for careful reasoning. On balance, we therefore believe that our results support proponents of quantifying probability assessments in national security, so long as those proponents do not expect that probabilistic precision is a free lunch: putting such measures in place calls for training and feedback that, while presumably feasible, demands rigorous analysis.

Our results suggest three other directions for further research. First, scholars can explore the mechanisms by which quantifying probability assessments reduces support for risky actions. In particular, it is important to understand whether this additional caution represents more careful considerations of risk as opposed to an illogical bias against basing decisions on numerical judgments.

Second, our empirical results call for further study of why respondents show consistent differences in qualitative versus quantitative assessments. This finding suggests that qualitative and quantitative probability assessments engage different cognitive processes. Further research can investigate what factors can explain and mitigate this divergence, while also exploring the generality of the pattern across contexts and cultures.

Third, and most broadly, we seek to reorient debates about expressing probability in national security from epistemology to empirics. Most published objections to quantifying subjective probabilities in national security revolve around normative claims about what kinds of language seem most appropriate for conveying the inherently subjective nature of national security analysis. For many scholars and practitioners, expressing subjective judgments using numbers simply feels wrong. But ultimately, if probabilistic precision threatens national security analysis and decision making, then that should have observable empirical consequences. If not, then intuitive discomfort with quantitative expression is a poor basis for sacrificing evidence-based improvements. This study offers evidence, consistent across both elite and general samples of respondents, sharpening claims about which consequences of quantifying subjective probabilities 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

Alpert, Marc, and Howard Raiffa. 1982. "A Progress Report on the Training of Probability Assesors" in Daniel Kahneman, Paul Slovic, and Amos Tversky eds., *Judgment Under Uncertainty*. New York: Cambridge 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 Broomell, and Michael Smithson. 2014. 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.

Connable, Ben. 2012. *Embracing the Fog of War*. Santa Monica, Calif.: RAND.

Dawes, Robyn M. 1988. *Rational Choice in an Uncertain World*. San Diego, Calif.: Harcourt Brace Jovanovich.

Dhami, Mandeep K., David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. Improving Intelligence Analysis with Decision Science. *Perspectives on Psychological Science* 10 (6): 743-757.

Friedman, Jeffrey A., and Richard Zeckhauser. 2012. Assessing Uncertainty in Intelligence. *Intelligence and National Security* 27 (6): 824-847.

------. 2015. Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden. *Intelligence and National Security* 30 (1): 77-99.

Friedman, Jeffrey A., Joshua D. Baker, Barbara A. Mellers, Philip E. Tetlock, and Richard Zeckhauser. 2015. The Value of Precision in Geopolitical Forecasting: Empirical Foundations for Intelligence Analysis and Foreign Policy Decision Making. Paper prepared for 2015 Annual Meeting of the American Political Science Association.

Hyde, Susan D. 2015. Experiments in International Relations: Lab, Survey, and Field. *Annual Review of Political Science* 18: 403-424.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. A General Approach to Causal Mediation Analysis. *Psychological Methods* 15 (4): 309-334.

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

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

Kent, Sherman. 1964. Words of Estimative Probability. *Studies in Intelligence*.

Kertzer, Joshua D. and Ryan Brutger. Forthcoming. Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory. *American Journal of Political Science*.

Krosnick, Jon A. and Leandre R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys" in L. Lyberg et al., *Survey Measurement and Process Quality*. New York: Wiley.

Lowenthal, Mark M. 1999. *Intelligence: From Secrets to Policy*. Washington, DC: CQ Press.

Mandel, David R. and Alan Barnes. 2014. Accuracy of Forecasts in Strategic Intelligence. *Proceedings of the National Academy of Sciences*. EarlyView.

Marchio, James. 2014. "If the Weatherman Can…": The Intelligence Community's Struggle to Express Analytic Uncertainty in the 1970s. *Studies in Intelligence* 58 (4): 31-42.

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

McDermott, Rose. 1998. *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy*. Ann Arbor, Mich.: University of Michigan Press.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, Pavel Atanasov, S. A. Swift, T. Murray, E. Stone, and Philip E. Tetlock. 2014. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science* 25 (5): 1106-15.

Moore, Don A. and Paul J. Healy. 2008. The Trouble with Overconfidence. *Psychological Review* 115 (2): 502-517.

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.

Powell Robert. 2002. Bargaining Theory and International Conflict. *Annual Reviews of Political Science* 5: 1-30.

Rapport, Aaron. 2015. *Waging War, Planning Peace: U.S. Noncombat Operations and Major Wars*. Ithaca, N.Y.: Cornell University Press.

Rathbun, Brian C. 2007. Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory. *International Studies Quarterly* 51 (3): 533-557.

Renshon, Jonathan. 2015. Losing Face and Sinking Costs: Experimental Evidence on the Judgment of Political and Military Leaders. *International Organization* 69 (3): 659-695.

Rieber, Steven. 2004. Intelligence Analysis and Judgmental Calibration. *International Journal of Intelligence and CounterIntelligence* 17 (1): 97-112.

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.

Sunstein, Cass R. 2007. *Worst-case Scenarios*. Cambridge, Mass.: Harvard University Press.

Tetlock, Philip E. 1999. Theory-Driven Reasoning about Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions? *American Journal of Political Science* 43 (2): 335-366.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, N.J.: Princeton University Press.

------ and Daniel Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. New York: Crown.

------ and Barbara A. Mellers. 2011. Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong. *American Psychologist* 66 (6): 542-554.

Tomz, Michael R. and Jessica P. Weeks. 2013. Public Opinion and the Democratic Peace. *American Political Science Review* 107 (4): 849-865.

Wallsten, Thomas, and David V. Budescu. 1995. A Review of Human Linguistic Probability Processing: General Principles and Empirical Evidence. *Knowledge Engineering Review* 10 43-62.

Weiss, Charles. 2008. Communicating Uncertainty in Intelligence and Other Professions," *International Journal of Intelligence and CounterIntelligence* 21 (1): 57-85.

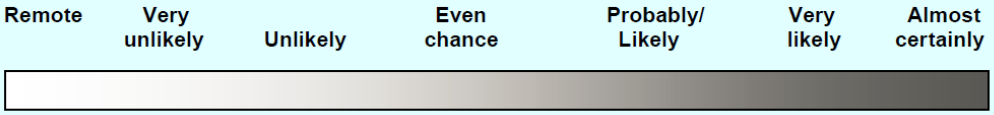
Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton, N.J.: Princeton University Press.

Zimmer, Alf C. 1994. A Model for the Interpretation of Verbal Predictions. *International Journal of Man-Machine Studies* 20: 121-134.

Figure 1. Examples of U.S. Intelligence Guidelines for Expressing Probability

"Words of Estimative Probability" appearing in National Intelligence Estimates since 2007

Estimates of Likelihood. Because analytical judgments are not certain, we use probabilistic language to reflect the Community's estimates of the likelihood of developments or events. Terms such as *probably*, *likely, very likely,* or *almost certainly* indicate a greater than even chance. The terms *unlikely* and *remote* indicate a less then even chance that an event will occur; they do not imply that an event will not occur. Terms such as *might* or *may* reflect situations in which we are unable to assess the likelihood, generally because relevant information is unavailable, sketchy, or fragmented. Terms such as *we cannot dismiss, we cannot rule out,* or *we cannot discount* reflect an unlikely, improbable, or remote event whose consequences are such that it warrants mentioning. The chart provides a rough idea of the relationship of some of these terms to each other.



"Words of Estimative Probability" recommended by the U.S. Director of National Intelligence in 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%

Figure 2. Examples of U.S. Military Guidelines for Expressing Probability

C. Guidelines for military risk assessment in U.S. Army Field Manual 5-19, "Composite Risk Management":

RISK ASSESSMENT MATRIX						
		Probability				
Severity		Frequent A	Likely B	Occasional C	Seldom D	Unlikely E
Catastrophic	I	E	E	H	H	M
Critical	II	E	H	H	M	L
Marginal	III	H	M	M	L	L
Negligible	IV	M	L	L	L	L
E – Extremely High		H – High		M – Moderate		L – Low

D. Guidelines for assessing human source reliability in U.S. Army Field Manual 2-22.3, "Human Intelligence Collector Operations":

Table B-1. Evaluation of Source Reliability.

A	**Reliable**	**No doubt** of authenticity, trustworthiness, or competency; has a history of complete reliability
B	**Usually Reliable**	**Minor doubt** about authenticity, trustworthiness, or competency; has a history of valid information most of the time
C	**Fairly Reliable**	**Doubt** of authenticity, trustworthiness, or competency but has provided valid information in the past
D	**Not Usually Reliable**	**Significant doubt** about authenticity, trustworthiness, or competency but has provided valid information in the past
E	**Unreliable**	**Lacking** in authenticity, trustworthiness, and competency; history of invalid information
F	**Cannot Be Judged**	**No basis** exists for evaluating the reliability of the source

Figure 3. Cumulative distributions of respondent Brier scores by treatment group

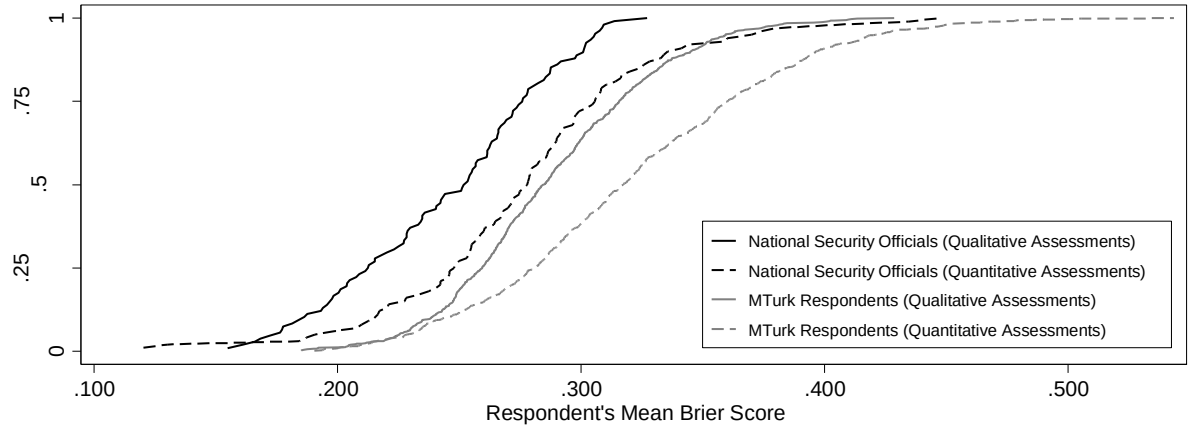


Table 1. Support for proposed actions across scenarios

Response measure: Scale of 1-7, with 1 defined as "Strongly oppose," 4 defined as "Neither support nor oppose," and 7 defined as "Strongly support."

			Scenario Version	
Scenario	***Sample***	**Pessimistic**	**Neutral**	**Optimistic**
All	*Natl Security Officials*	2.72 (1.54)	3.74 (1.86)	4.51 (1.92)
	MTurk	2.93 (1.69)	3.75 (1.82)	4.53 (1.76)
Hostage	*Natl Security Officials*	3.53 (1.91)	4.83 (1.73)	5.63 (1.52)
	MTurk	3.78 (1.71)	4.72 (1.71)	5.39 (1.44)
Drone	*Natl Security Officials*	2.21 (1.08)	2.84 (1.77)	3.38 (1.87)
	MTurk	2.78 (1.73)	3.49 (1.77)	4.12 (1.85)
Security	*Natl Security Officials*	2.44 (1.18)	3.53 (1.53)	4.52 (1.66)
	MTurk	2.23 (1.21)	3.06 (1.54)	4.09 (1.63)

Table 1 presents mean (standard deviation) support for approving proposed actions across scenarios, in both elite and general samples. Differences in means across scenarios versions are all statistically significant at the $p<0.001$ level. See appendix for question and scenario wordings.

Table 2a. Responses to Scenarios – National Security Officials

	Model 1: Predicting support for risky action	*Model 2:* Predicting support for risky action, with interaction terms	*Model 3:* Predicting support for delaying action	*Model 4:* Predicting confidence levels
Quantitative assessment	-0.142 (.14)	-0.346 (.25)	0.316 (.15)*	-0.168 (.11)
Optimistic scenario	0.783 (.16)***	0.677 (.23)**	-0.656 (.18)***	0.134 (.09)
Pessimistic scenario	-0.982 (.16)***	-1.146 (.22)***	0.424 (.16)**	0.275 (.09)**
Hostage scenario	1.148 (.16)***	1.152 (.16)***	-0.227 (.18)	0.170 (.08)*
Drone scenario	-0.617 (.14)***	-0.619 (.14)***	1.120 (.16)***	0.317 (.08)***
Numeracy	-0.157 (.06)**	-0.155 (.06)**	0.102 (.06)	-0.035 (.05)
Female	-0.257 (.21)	-0.261 (.21)	0.129 (.22)	-0.136 (.19)
Military officer	0.185 (.17)	0.181 (.17)	-0.187 (.16)	0.223 (.15)
U.S. citizen	0.096 (.35)	0.078 (.34)	-0.132 (.36)	0.687 (.28)*
English as native lang.	0.143 (.34)	0.145 (.34)	-0.102 (.37)	-0.872 (.26)***
Optimistic scenario x Quantitative assessment		0.232 (.33)		
Pessimistic scenario x Quantitative assessment		0.364 (.31)		
Constant	3.739 (.29)***	3.841 (.31)***	4.58 (.30)***	5.304 (.25)***
Overall R^2	0.298	0.300	0.165	0.059

Ordinary least squares regressions predicting 7-point response measures with respondent fixed effects. Robust standard errors.
* p<0.05 ** p<0.01 *** p<0.001. All models have 624 observations over 208 respondents.

Table 2b. Responses to Scenarios – MTurk Respondents

	Model 1: Predicting support for risky action	*Model 2:* Predicting support for risky action, with interaction terms	*Model 3:* Predicting support for delaying action	*Model 4:* Predicting confidence levels
Quantitative assessment	-0.428 (.05)***	-0.759 (.09)***	0.382 (.06)***	0.098 (.06)
Optimistic scenario	0.795 (.06)***	0.481 (.08)***	-0.523 (.07)***	0.050 (.04)
Pessimistic scenario	-0.811 (.06)***	-0.970 (.09)***	0.216 (.07)***	0.148 (.05)***
Hostage scenario	1.500 (.06)***	1.505 (.06)***	-0.329 (.06)***	0.048 (.04)
Drone scenario	0.328 (.06)***	0.327 (.06)***	0.451 (.06)***	0.254 (.04)***
Numeracy	-0.113 (.02)***	-0.112 (.02)***	0.048 (.03)	-0.036 (.02)
Female	0.006 (.06)	-0.002 (.06)	0.265 (.06)***	-0.248 (.06)***
Military service	0.163 (.13)	0.156 (.13)	-0.385 (.15)*	0.209 (.13)
U.S. citizen	0.073 (.27)	0.065 (.27)	-0.313 (.26)	-0.206 (.28)
English as native lang.	-0.259 (.24)	-0.252 (.24)	-0.345 (.25)	-0.176 (.22)
Education	0.069 (.04)	0.070 (.04)	$2.9e^{-4}$ ($3.0e^{-3}$)	-0.056 (.04)
Age	0.001 (.00)	0.001 (.00)	0.003 (.00)	0.013 (.00)***
Optimistic scenario x Quantitative assessment		0.654 (.12)***		
Pessimistic scenario x Quantitative assessment		0.324 (.12)**		
Constant	3.599 (.35)***	3.759 (.36)***	5.211 (.31)***	5.160 (.32)***
Overall R^2	0.260	0.265	0.078	0.030

Ordinary least squares regressions predicting 7-point response measures with respondent fixed effects. Robust standard errors.
* p<0.05 ** p<0.01 *** p<0.001. All models have 4,376 observations over 1,459 respondents.

Table 3. Predictors of Brier Scores

	National Security Officials	Amazon Mechanical Turk
Quantitative Assessment	0.031 (.01)***	0.032 (.00)***
Numeracy	-0.002 (.00)	-0.004 (.00)***
Military Service	0.001 (.01)	0.006 (.02)
Female	0.007 (.01)	0.002 (.00)
U.S. Citizen	0.017 (.02)	0.003 (.01)
English Native Lang.	-0.035 (.02)	0.002 (.01)
Education Level		-0.011 (.00)***
Constant	0.265 (.01)***	0.310 (.01)***
N	208 respondents	1,561 respondents
R^2	0.140	0.126

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 4. Comparing Distributions of Qualitative/Quantitative Probability Assessments

	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)
				National Security Officials			
Qualitative assessments	4.51	7.10	16.67	12.87	25.28	20.34	13.24
Quantitative assessments	15.73	8.73	8.60	15.60	8.87	16.33	26.13
	p<0.001	*p*=0.017	*p*<0.001	*p*=0.002	*p*<0.001	*p*<0.001	*p*<0.001
				MTurk Respondents			
Qualitative assessments	1.65	4.94	15.66	19.75	29.13	19.30	9.57
Quantitative assessments	10.90	9.98	10.64	18.12	13.66	16.12	20.58
	p<0.001	*p*<0.001	*p*<0.001	*p*<0.001	*p*<0.001	*p*<0.001	*p*<0.001

Table 3 compares the proportion of estimates falling within each segment of the number line, registered by respondents in the qualitative vs. quantitative assessment conditions. Statistical significance estimated via two-way t-tests.

Table 5. Causal mediation analysis of quantitative assessment, respondent certitude, and accuracy

OLS regression predicting respondent certitude

	National Security Officials	Amazon Mechanical Turk
Quantitative Assessment	0.043 (.00)***	0.040 (.00)***
Numeracy	-0.002 (.01)	-0.001 (.00)
Military Service	0.003 (.01)	0.030 (.01)*
Female	-0.000 (.01)	-0.007 (.00)***
U.S. Citizen	0.031 (.01)	-0.007 (.01)
English Native Lang.	-0.029 (.01)*	0.005 (.01)
Birth Year		$-1.7e^{-5}$ ($1.1e^{-5}$)
Education Level		$-1.0e^{-4}$ ($5.9e^{-4}$)
Constant	0.819 (.01)	0.846 (.02)***
N	208 respondents	1,558 respondents
R^2	0.332	0.347

*p<.05, **p<.01, ***p<.001

Panel B: OLS regression predicting Brier Scores

	National Security Officials	Amazon Mechanical Turk
Quantitative Assessment	0.025 (.01)**	0.006 (.00)*
Certitude	0.140 (.10)	0.650 (.04)***
Numeracy	-0.002 (.00)	-0.004 (.00)***
Military Service	0.001 (.01)	-0.001 (.02)
Female	0.007 (.01)	0.007 (.00)**
U.S. Citizen	0.012 (.02)	0.007 (.01)
English Native Lang.	-0.031 (.02)	-0.002 (.01)
Birth Year		$-1.6e^{-5}$ ($1.8e^{-5}$)
Education Level		-0.006 (.00)***
Constant	0.150 (.09)	-0.240 (.05)
N	208 respondents	1,558 respondents
R^2	0.148	0.249

Note that lower Brier Scores indicate more accurate assessments. *p<.05, **p<.01, ***p<.001

Appendix. Scenario Text

We present the text of each vignette in its "neutral" version. See supplementary material for more information.

Scenario 1

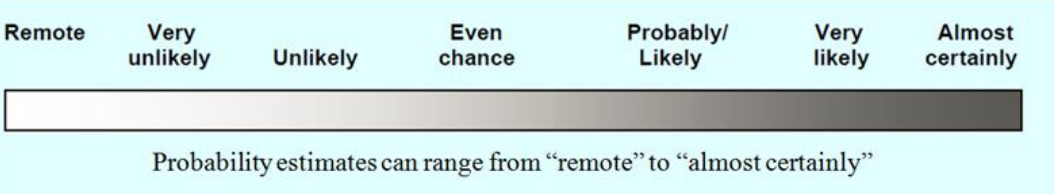
ISIS is holding three American aid workers hostage. The U.S. Intelligence Community has used human intelligence and communications intercepts to trace these hostages to a compound in Manbij, Syria.

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 / 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, [it is very likely / there is an 80 percent chance] that Special Forces can retrieve the hostages alive.

This mission entails several risks. Analysts believe there is [an even chance / a 50 percent chance] that ISIS will wound or kill U.S. soldiers on this mission. They believe that [it is possible, though unlikely, / 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), [ISIS will almost certainly / there is a 95 percent chance that ISIS will] execute the hostages.

Summary of estimated chances:

- The hostages are at the Manbij compound: [*Likely* / *65 percent*]
- If the hostages are in this location, Special Forces can retrieve them alive: [*Very likely* / *80 percent*]
- ISIS will wound or kill U.S. soldiers on this mission: [*Even chance* / *50 percent*]
- The mission would inadvertently wound or kill innocent civilians: [*Unlikely* / *35 percent*]
- ISIS will kill the hostages if the raid fails: [*Almost certainly* / *95 percent*]



[This spectrum displayed in the qualitative assessment condition only]

Scenario 2

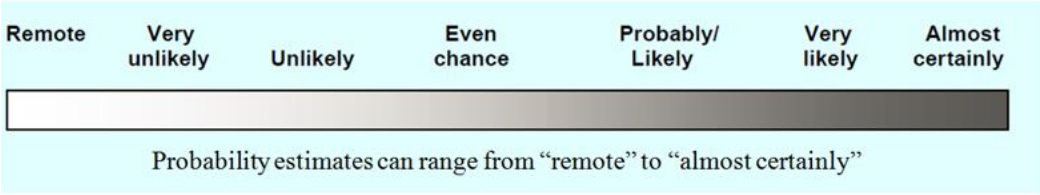
The Central Intelligence Agency uses drones to monitor houses in Yemen believed to be used by Al Qaeda in the Arabian Peninsula (AQAP). CIA analysts report that an unusual number of people have recently been gathering at one of these houses. At any given time, there are at least 8-12 individuals inside 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 / there is a 65 percent chance] 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 / there is an 80 percent chance] that a drone strike on the house would kill everyone inside.

Analysts warn that [it is possible, though unlikely, / there is a 35 percent chance] that the house contains women and children. If U.S. forces strike this target, then [it is almost certain / there is a 95 percent chance] 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* / *65 percent*]
- A drone strike on the house would kill everyone inside: [*Very likely* / *80 percent*]
- The house contains women and children: [*Unlikely* / *35 percent*]
- The drone strike will compromise ongoing surveillance efforts in the area: [*Almost certainly* / *95 percent*]



[This spectrum displayed in the qualitative assessment condition only]

Scenario 3

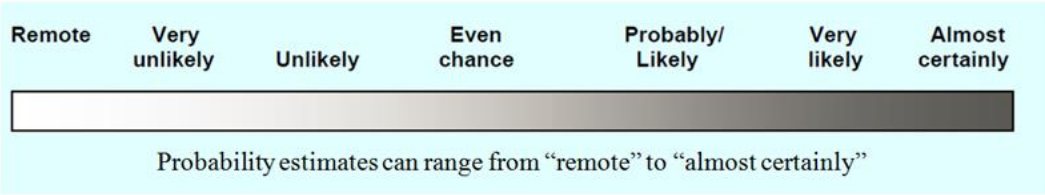
An Afghan leader named Ghamay Jan recently approached U.S. officials. Jan offered to mobilize 500 followers to combat the Taliban along a dangerous stretch of the border with Pakistan located in Khost Province. Jan requests that the United States provide him funding, equipment, and permission to use 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 / there is a 65 percent chance] 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 / there is a 65 percent chance] 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 / a 50 percent chance] that he previously assisted the Taliban to establish a presence in this part of Khost Province. They also say [it is very likely / there is an 80 percent chance] that Jan would use the authority he requests to facilitate illegal smuggling. If the United States supports Ghamay Jan, analysts say [it is unlikely / there is a 35 percent chance] 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* / *65 percent*]
- If Jan's forces cooperate with the United States, they would prevent Taliban from crossing the nearby border: [*Likely* / *65 percent*]
- Jan previously assisted the Taliban: [*Even chance* / *50 percent*]
- Jan will use the authority he requests to secure illegal smuggling: [*Very likely* / *80 percent*]
- The United States can retain backing from other local leaders if U.S. forces support Ghamay Jan: [*Unlikely* / *35 percent*]



[This spectrum displayed in the qualitative assessment condition only]