Analytic Confidence and Political Decision Making: Theoretical Principles and Experimental Evidence from National Security Professionals

Jeffrey A. Friedman, Assistant Professor of Government, Dartmouth College
Richard Zeckhauser, Frank P. Ramsey Professor of Political Economy, Harvard University

Working Paper: May 2017

Contact: jeffrey.a.friedman@dartmouth.edu

Abstract. When making decisions under uncertainty, it is important to distinguish between the probability that a judgment is true and the confidence analysts possess in drawing their conclusions. Yet analysts and decision makers often struggle to define “confidence” in this context, and many ways that scholars use this term do not necessarily facilitate decision making under uncertainty. To help resolve this confusion, we argue for disaggregating analytic confidence into three dimensions: reliability of available evidence, range of reasonable opinion, and responsiveness to new information. After explaining how these attributes hold different implications for decision making in principle, we present survey experiments examining how analysts and decision makers employ these ideas in practice. Our first experiment found that each conception of confidence distinctively influenced national security professionals’ evaluations of high-stakes decisions. Our second experiment showed that inexperienced assessors of uncertainty could consistently discriminate among our conceptions of confidence when making political forecasts. We focus on national security, where debates about defining “confidence levels” have clear practical implications. But our theoretical framework generalizes to nearly any area of political decision making, and our empirical results provide encouraging evidence that analysts and decision makers can engage these abstract elements of uncertainty.

This research was supported by the United States Department of Homeland Security (DHS) through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) at the University of Southern California (USC) under award number 2010-ST-061-RE0001. Geoffrey Berlin, Mark Bucknam, Stephen Dyson, Thomas Fingar, Paul Heer, Yusaku Horiuchi, Dean Lacy, Mark Lowenthal, Brendan Nyhan, Gregory Treverton, Kristan Wheaton, and Laura Yost provided helpful comments on previous drafts, as did seminars at the National Intelligence Council, the Department of Homeland Security’s Science and Technology Directorate, the 84th Military Operations Research Society Symposium, the Ridge College of Intelligence Studies and Applied Sciences at Mercyhurst University, and the 2017 annual meeting of the International Studies Association. Freya Jamison and Ben Rutan provided outstanding research assistance. We are especially grateful to the 223 national security officials who volunteered their time to participate in our survey experiments. Any opinions, findings, conclusions, or recommendations in this document are those of the authors and do not necessarily reflect views of DHS, USC, or CREATE.
Analytic Confidence and Political Decision Making

This paper explores the distinction between probability and confidence in political decision making. Broadly speaking, probability reflects an analyst’s estimate of the chances that a statement is true, while confidence reflects the degree to which an analyst believes that he or she possesses a sound basis for assessing uncertainty. In U.S. criminal courts, for example, jurors should not send defendants to jail just because their guilt seems likely. In principle, guilty verdicts require concrete evidence, not just subjective certainty. The criminal justice system thus requires jurors to assess both high probability and high confidence to find for conviction.

The relationship between probability and confidence shapes many political issues. Scholars often argue that the use of force in international relations, like the removal of civil liberties in a court of law, must be justified through reliable evidence and not merely the belief that an action will produce favorable results. When scholars and pundits debate election forecasts, they are (at least implicitly) arguing both about what probabilities they should assign to electoral outcomes and about how much confidence those estimates deserve. Though virtually all scientists believe that carbon emissions are harming the global climate, skeptics observe that scientists lack a commonly-accepted framework for modeling climate change, and therefore argue that these predictions are insufficiently reliable to justify costly action. In this way, political judgment requires assigning probabilities to significant events as well as judging the quality of those estimates before their accuracy can be known. This is the sense in which we use the term “confidence” (or “analytic confidence”) in this paper.

Political scientists have said relatively little about analytic confidence in this context, even as they have developed a robust literature on the nature of “good judgment” more generally
Friedman and Zeckhauser, “Analytic Confidence”

(Renshon and Larson, 2003; Schafer and Crichlow, 2010; Tetlock, 1992). Indeed, most of the ways in which the term “confidence” appears in contemporary scholarship have little to do with judging the extent to which analysts believe they possess a sound basis for making a particular assessment of uncertainty. Some scholars define confidence as an attribute belonging to individuals, such as how precisely analysts tend to estimate uncertain quantities, or how highly they rate their overall performance (Moore & Healy, 2008; Ortoleva & Snowberg, 2015). Others define confidence as an attribute of available data, such as how “confidence intervals” characterize statistical parameters. Still other scholars define confidence based on ex post evaluations of judgmental accuracy, such that an analyst is “overconfident” if she assigns excessive certainty to her judgments (Johnson, 2004; Tetlock, 2005). When survey researchers ask respondents to say how “confident” they are in their political knowledge, this is often equivalent to eliciting beliefs about the chances that those statements are true.

The fact that the word “confidence” conveys so many distinct ideas provides a pragmatic reason to employ clearer terminology when debating high-stakes issues. But this paper also provides a theoretical argument for why traditional discussions of analytic confidence are excessively vague. In particular, we explain how there are at least three distinct ways of describing the extent to which an analyst believes that he or she possesses a sound basis for assessing uncertainty: reliability of available evidence, range of reasonable opinion, and responsiveness to new information. We distinguish these attributes from probability estimates, explain how they can vary independently, and argue that each carries different implications for political decision making. We focus this discussion on national security, as this is a domain where scholars have long debated conceptual frameworks for opening up the “black box” of uncertainty (Rathbun, 2007; Vertzberger, 1995), and where these debates directly shape the
production of intelligence reports, military plans, and other impactful analyses (Fingar, 2011; Nye, 1994). Yet we emphasize throughout the paper how our three conceptions of confidence apply to nearly any other area of political decision making where uncertainty plays a role.

After developing a theoretical framework for understanding the relationship between analytic confidence and political decision making in principle, we present two survey experiments examining the extent to which respondents can employ these ideas in practice. Our first survey experiment shows how a cross-section of national security elites, recruited through the National War College, employed all three conceptions of confidence simultaneously when evaluating decisions under uncertainty. Our second survey experiment demonstrated that even novice analysts, recruited from Amazon Mechanical Turk, could consistently discriminate among our three conceptions of confidence when making political forecasts.

Beyond supporting the conceptual validity of our framework for decomposing analytic confidence into three separate attributes, our findings speak to broader academic questions about the ability of political analysts and decision makers to handle abstract ideas when assessing uncertainty. Behavioral studies often find that respondents’ perceptions of uncertainty are more consistent with affect-laden feelings than with reasoned judgments (Gigerenzer, 2008; Kahneman, 2011, 201; Loewenstein, Hsee, Weber, & Welch, 2001; Slovic, 2010). This skepticism is particularly strong in national security, where many scholars view the challenge of assessing uncertainty as being unusually complex and subjective, and therefore unsuited to analytic precision (Betts, 2000; Jervis, 1997). This paper recognizes that perceptions of uncertainty are subjective and susceptible to bias. Nevertheless, we demonstrate that both elite and non-elite respondents can consistently distinguish between probability and confidence, and that they can reliably divide assessments of confidence into three constituent parts. Our findings
thus bolster efforts to improve assessments of uncertainty in public discourse (Ho, Budescu, Dhami, and Mandel 2015), while providing encouraging evidence that analysts and decision makers can engage with abstract elements of uncertainty.

Three conceptions of confidence

The challenge of disentangling probability and confidence is especially prominent in the domain of intelligence and national security. Figure 1 presents current guidelines on this subject from the U.S. Defense Intelligence Agency (DIA) and National Intelligence Council (NIC). Though these standards instruct analysts to convey probability estimates using words instead of numbers, there is little doubt that these phrases correspond to ranges of numbers between zero percent and one hundred percent. By contrast, these guidelines define confidence with respect to attributes such as quality and quantity of available information, potential for deception, gaps in knowledge, and strength of relevant inferences. It is unclear how analysts should operationalize any of these attributes, let alone how one might express those factors’ combined magnitude on a single scale. Vagueness surrounding confidence levels thus extends beyond ambiguous language: interpreting these assessments requires grappling with conceptual confusion about what analytic confidence means.

[Figure 1]

This confusion has two main elements. First, it is unclear why some factors associated with analytic confidence should matter for decision making. For example, both the DIA and the NIC define confidence with respect to the number of sources that analysts use to draw conclusions. But if an analyst finds a single, dispositive source, then there is no reason why she cannot make a
sound judgment, just as one credible eye-witness sometimes justifies a criminal conviction. Thus one challenge in defining analytic confidence is understanding which factors represent primary elements of uncertainty, and which proxy for other, more important concepts. In this respect, our theoretical framework hones discussions of confidence into three basic ways of describing the extent to which analysts believe they possess a sound basis for assessing uncertainty. We call these attributes reliability of available evidence, range of reasonable opinion, and responsiveness to new information.

Existing discussions of analytic confidence also tend to be underspecified. For example, when the DIA describes “strength of analysis” or when the NIC discusses “high-quality information,” these phrases suggest constellations of factors as opposed to precise ideas. In describing our three conceptions of confidence, we thus explain how each can be quantified in principle. This is not to claim that analysts should quantify these attributes in practice, but rather to demonstrate that our conceptions of confidence reflect distinct and coherent elements of assessing uncertainty. In conducting this discussion, we often rely on the concept of “second-order uncertainty,” which captures the relative credibility that analysts assign to every possible judgment. Second-order uncertainty thus reflects an analyst’s belief about the chances that the “right” probability estimate to make on the basis of available evidence is zero percent, one percent, two percent, and so on up to one hundred percent. Each of our three conceptions of confidence articulates a different way in which second-order uncertainty shapes political decision making in a manner that probability estimates alone cannot capture.

Of course, we cannot say that reliability of available evidence, range of reasonable opinion, and responsiveness to new information are the only ways of describing analytic confidence that would ever matter to making political decisions. Together, however, our three conceptions of
confidence cover each basic element of Bayesian reasoning. Reliability of available evidence reflects the degree to which an analyst’s judgments depend on her prior assumptions; range of reasonable opinion captures an analyst’s current perception of second-order uncertainty; responsiveness to new information reflects how further study might shift those perceptions into what decision theorists call “posterior” judgments. At the very least, we will argue that our three conceptions of confidence should matter for making all political decisions where uncertainty plays a role. To our knowledge, our theoretical framework reflects the first attempt to specify generalizable components of analytic confidence in this manner.

Reliability of available evidence

Our first conception of analytic confidence is reliability of available evidence. In this view, analysts have a sound basis for assessing uncertainty when they base their judgments on credible information as opposed to speculation. This is the most common way that scholars and practitioners define confidence levels in intelligence studies. Thus Thomas Fingar, formerly Deputy Director of National Intelligence for Analysis, writes that “confidence judgments are based on the quantity, quality, and consistency of the information available” (Fingar, 2011, 88; cf. Wheaton, 2012, 335). Assessing the reliability of available evidence requires considering the volume of available information, the diagnostic value of individual data points, analysts’ subject matter expertise, and the extent to which independent sources provide corroborating views. We
can summarize the combined relevance of these factors with respect to a single attribute, which is the extent to which an analyst’s judgment depends on her prior beliefs.¹

While judgments based on more reliable evidence often entail greater certainty, those attributes vary independently. For example, estimates of the chances that a card drawn from a randomly-shuffled deck will be black (one-in-two), a diamond (one-in-four) or an ace (one-in-thirteen) all reflect equally-reliable reasoning. Identical probability estimates can also reflect evidence of varying reliability. Indeed, one of the main reasons why governments establish intelligence agencies, or other analytic organizations like the Council of Economic Advisers and the Congressional Budget Office, is to help decision makers understand which assessments of uncertainty are reliable and which are not.

In classical decision theory, the reliability of available evidence does not directly shape decision making. Consider a gamble where you can flip either of two coins. You win a prize if you guess the coin flip’s outcome correctly. One is a fair coin with equal chances of landing heads versus tails. The other coin is arbitrarily weighted: its chances of landing heads could be anywhere from zero to one hundred percent, and you have no information about what this probability might be. These gambles offer identical expected utility. In the first case you know that the probability of the coin landing heads is fifty percent. In the second case, you have no

¹ We can quantify this attribute as the statistical difference between an analyst’s second-order uncertainty and what that judgment would have been were it based on uninformed prior assumptions that assigned equal credibility to every hypothesis. This captures the degree to which the analyst’s prior assumptions continue to shape her assessment of uncertainty after viewing the available evidence.
reason to suspect that either outcome is more likely than the other. Your expected probability of
the coin landing heads thus remains fifty percent. (If you called the outcome of this coin flip at
random, that alone would guarantee you a fifty percent chance of success.) Though many
individuals would not treat these gambles equally, and might even pay nontrivial amounts in
order to avoid flipping the weighted coin, most decision theorists see this “ambiguity aversion”
as misguided (Al-Najjar & Weinstein, 2009).

Reliability of available evidence nevertheless carries important implications for the ethics of
political decision making. In particular, many people believe that it is inappropriate for public
policy to cause harm without concrete justification. Controversy over racial profiling by police is
thus largely a debate over whether it is acceptable to impose costs on individual citizens without
indicators of personal guilt. Similarly, when debating environmental regulations, drug approvals,
and the proposed privatization of social security, decision makers must consider the extent to
which they can prove that their actions do not impose unnecessary risks. In this sense, reliability
of available evidence shapes the ethics – and thereby the politics – of public policy in a manner
that cannot be expressed simply through estimated probabilities or the utilities that decision
makers attach to policy outcomes.

George W. Bush’s presidency offers two demonstrations of how this dynamic shapes national
security decision making. The first case is the 2003 invasion of Iraq. The Bush administration
publicly justified the war as a measure to preempt Saddam Hussein from obtaining nuclear
weapons. That claim relied on circumstantial evidence interpreted in light of Saddam’s past
behavior. Many scholars believe that the speculative nature of this analysis undermined the Bush
administration’s appeal to preemption, as invoking that doctrine requires clear evidence of
imminent threat (Doyle, 2008; Walzer, 2006). In this view, one of the main flaws with pre-war
intelligence on Iraq was not that analysts concluded Saddam was likely pursuing nuclear weapons, but that published reports failed to clarify the degree to which this conclusion relied on prior beliefs (Betts, 2007, 116; Jervis, 2006, 44). According to former Deputy Director of Intelligence Michael Morell, “By far the biggest mistake made by the analysts… was not that they came to the wrong conclusion about Iraq’s WMD program, but rather that they did not rigorously ask themselves how confident they were in their judgments” (Morell, 2015, 102).

A similar dynamic appeared later in Bush’s presidency in response to reports that Syria was building a nuclear reactor at a site called al-Kibar. Former CIA Director Michael Hayden (2016, 262) describes briefing President Bush about this subject in 2007, explaining that while analysts were virtually certain that al-Kibar was a military facility, they possessed no reliable evidence supporting this claim. “This is part of a Syrian weapons program. Of course it is,” Hayden recalls saying. Yet he continued, “I can’t find the other parts of the weapons program. No reprocessing facility. No weaponization effort that we can see. So I can only give this to you with low confidence.” Hayden then recounts how even though some of President Bush’s top advisers recommended destroying the facility with an air strike, “the president observed that his preemption policy demanded a threat be imminent before we could act. Our estimate of low confidence in a weapons program made that very difficult to justify, and therefore, the president declared, ‘We will not strike the facility.’” This statement highlights how subjective probability estimates alone do not capture how assessments of uncertainty inform high-stakes political decisions. In both the Iraq and al-Kibar cases, the reliability of available evidence also shaped the ethical and political context for taking action.
Range of reasonable opinion

Our second conception of confidence is range of reasonable opinion. In this view, analysts have a sounder basis for assessing uncertainty when they believe the set of plausible viewpoints is narrower. Range of reasonable opinion thus resembles the way that social scientists describe “confidence intervals,” though given how most statistical models rely on debatable assumptions, it is usually a mistake to believe that statistical output objectively characterizes the uncertainty surrounding a given inference (Mankisi 2013). Similarly, an analyst can gauge variations in judgment by surveying experts, but she might believe that some of these responses are unreasonable or that some opinions falling outside this range are reasonable. Analysts can always quantify the way they perceive the range of reasonable opinion by describing their second-order uncertainty as to what the right probability estimate entails.

In statistics, larger volumes of information often generate tighter confidence intervals. Nevertheless, reliability of available evidence and range of reasonable opinion vary independently. For example, econometric projections of stock prices typically entail huge uncertainty. Thus even when financial analysts base market forecasts solely on extensive data sets, they may still have to recognize that there is a wide range of reasonable opinion about a stock’s future price. Similarly, climate researchers who base their forecasts on scientific models must accept that different modeling assumptions generate substantial differences in predicted outcomes. The Intergovernmental Panel on Climate Change (IPCC) thus defines “confidence” based on reliability of available evidence and on range of reasonable opinion. As shown in Figure 2, IPCC guidelines state that assessments must reflect high degrees of both attributes in order to qualify as “high confidence” (Mastrandrea et al., 2010, 3).
Range of reasonable opinion can shape political decisions in three main ways. First – as with reliability of available evidence – this conception of confidence raises ethical concerns for taking action under conditions of ambiguity. In most U.S. criminal trials, juries can only convict defendants by unanimous vote. A criminal trial’s outcome thus depends on assessing probability (guilt beyond “reasonable doubt”), reliability of evidence (such as the exclusion of hearsay), and range of reasonable opinion (via jury consensus). The broader practice of establishing veto power in political decision making serves a similar purpose, and even when unanimity is not formally required to approve a policy, many organizations strive to achieve consensus in order to maximize the legitimacy of their actions.

The Iraq War again shows how this concern shapes the politics of national security decision making. The 2002 National Intelligence Estimate on Iraq’s weapons of mass destruction programs contained a dissenting opinion from the U.S. Department of State. State Department analysts disputed the claim that Iraq’s acquisition of aluminum tubes indicated its intent to build uranium centrifuges. Scholars disagree over how much weight this dissent deserved: Robert Jervis (2006, 18), for example, points out that “apparently all intelligence services in all countries and most private analysts came to roughly the same conclusions about the Iraqi programs.” The State Department’s opinion in this case was thus an outlier, and not all outliers are credible. Nevertheless, many observers argue that it was inappropriate for the U.S. Intelligence Community to offer such a high-stakes judgment, or for the Bush administration to act on that judgment, without conclusively rebutting the State Department’s dissent (U.S. Senate, 2005, 87-118). If nothing else, the existence of such dissenting opinions allows critics to claim
that analysts and decision makers should have understood the flaws in their conclusions *ex ante* (Rovner, 2011).

Describing the range of reasonable opinion can also be productive in prompting decision makers to explore the basis for existing disagreement. During the search for Osama bin Laden, for example, President Obama asked his advisers to quantify the chances that al Qaeda’s leader was living in Abbottabad, Pakistan. The head of the Central Intelligence Agency’s bin Laden team placed these chances at ninety-five percent, while CIA Deputy Director Michael Morell put his own guess at sixty percent. When President Obama asked about this discrepancy, Morell explained that while counterterrorism officials trusted their targeting abilities given recent tactical successes, his experience assessing Iraq’s weapons of mass destruction programs a decade earlier left him wary of basing strategic decisions on circumstantial evidence. Morell (2015, 156-161) recounts how this discussion played an important role framing President Obama’s choice over whether to strike the Abbottabad compound. For similar reasons, Richard Neustadt and Ernest May recommend that decision makers ask analysts to quantify probability estimates when debating major decisions, as “We know of no better way to force clarification of meanings while exposing hidden differences” (Neustadt & May, 1986, 152; for broader discussion on the productive uses of competing viewpoints for decision making, see Mercier & Landemore, 2012; Mintz & Wayne, 2016; Schafer & Crichlow, 2010).

*Responsiveness to new information*

Our third conception of confidence is responsiveness to new information. In this view, individuals should be more “confident” when assessing uncertainty the less they expect
subsequent analysis to change their judgments. Responsiveness to new information depends on several factors, including how much information analysts currently possess, how strongly they believe in their current judgments, and what additional information they might be able to collect if given additional time and resources (Friedman and Zeckhauser 2015). Formally, we can express these factors’ combined influence in terms of second-order uncertainty about the value a probability estimate might take at some future date.

Neither reliability of available evidence nor range of reasonable opinion addresses this subject directly. For example, some of the most important questions in political discourse are also the most subjective. What is the risk of another global financial crisis within the next decade? What are the chances that the next Supreme Court will overturn Roe v. Wade? What is the probability that the United States and China will go to war by 2030? Any answer to these questions will rely heavily on speculation while leaving wide room for reasonable disagreement. Many political analysts are reluctant to offer such judgments, believing that their opinions would lack sufficient confidence to inform decision making. Yet seen through the lens of responsiveness to new information, these questions often provide the strongest basis for assessing uncertainty, given that further study is unlikely to change analysts’ views within a relevant time frame.

Responsiveness to new information is particularly important for timing major decisions. Under conditions of uncertainty, decision makers can usually delay action in order to gather additional information. Yet gathering that information can be expensive, both in terms of the direct costs of analysis and the opportunity costs of delay. For example, when intelligence

---

2 This is particularly problematic given many political decision makers’ tendencies towards premature cognitive closure (Bar-Joseph & Kruglanski, 2003; Bar-Joseph & McDermott, 2017).
officials first approached President Obama with the idea that Osama bin Laden might be living in Abbottabad in fall 2010, Obama asked them to develop alternative methods for collecting information about the compound. Though some of these methods yielded valuable information, none conclusively identified the compound’s occupants (Morell 2015, 152). By spring 2011, it was clear that intelligence officials had exhausted feasible avenues for collecting actionable evidence. Meanwhile, the longer President Obama waited to authorize a raid on the compound, the greater the chances became that the people living in this compound would learn they were being watched, or that they would leave for other reasons. Thus regardless of the reliability of available evidence about the Abbottabad compound or the range of reasonable opinion surrounding this subject, President Obama determined that he could not justify further delay.

President Bush’s decision not to attack Syria’s nuclear reactor in 2007 also depended on weighing responsiveness to new information. Though all analysts agreed that the al-Kibar facility was likely designed for military purposes, they also believed that the reactor was years away from producing weapons-grade material. There was thus little opportunity cost in continuing to study the facility. Meanwhile, intelligence officials had viable opportunities to continue gathering information about the site, and it was possible that the situation would be resolved without U.S. intervention. (Indeed, Israel destroyed the facility with an air strike later that year.) Compared to the bin Laden case, analysts had higher certainty, more reliable evidence and greater analytic consensus in assessing al-Kibar. Yet President Bush chose not to strike the Syrian reactor, in part because gathering additional information was more attractive than striking immediately.
Friedman and Zeckhauser, “Analytic Confidence”

**Testing conceptual validity through survey experiments**

The previous section argued that the common practice of ascribing “confidence levels” to assessments of uncertainty obscures information that is relevant to making high-stakes decisions. We argued that analytic confidence is not a single variable, but rather three distinct attributes that vary independently, each with different implications for political decision making.

Yet explaining how our three conceptions of confidence operate in principle does not guarantee that these distinctions matter in practice. Behavioral researchers often find that analysts and decision makers struggle to manage uncertainty in rational ways. In many areas of high-stakes decision making, perceptions of uncertainty seem more consistent with affect-laden feelings than with reasoned judgments (Gigerenzer, 2008; Kahneman, 2011, 201; Loewenstein et al., 2001; Slovic, 2010). This presents two main challenges to the conceptual validity of our framework. First, decision makers may not reliably respond to distinctions among our three conceptions of confidence, in which case our ideas would add little descriptive value for understanding how political decisions are made. Second, analysts might be unable to discriminate among our three conceptions of confidence, implying that our ideas hold little prescriptive value for improving assessments of uncertainty in national security or other fields.

We conducted two survey experiments to address these concerns. Our first survey experiment shows how a cross-section of national security elites, recruited through the National War College, employed all three conceptions of confidence simultaneously when evaluating decisions under uncertainty. Our second survey experiment demonstrates that novice political analysts, recruited through Amazon’s Mechanical Turk survey platform, could consistently discriminate among three conceptions of confidence when making political forecasts. These findings suggest
that our attempts to open up the “black box” of analytic confidence have both descriptive and prescriptive value: we see how these distinctions shape the way that real national security officials evaluate high-stakes decisions, and we find that analysts can draw these distinctions intuitively.

Survey experiments have three main drawbacks for the study of political decision making. First, survey experiments cannot replicate the dense mixture of cues and pressures that real high-stakes decisions entail. However, the experimental method provides the clearest methodology for exploring how our three conceptions of confidence shape analysis and decision. As the previous section explained, analysts and decision makers often struggle to distinguish between probability and confidence. Even when analysts and decision makers describe analytic confidence in practice, they often leave these judgments vague and they never vary these assessments at random. Survey experiments thus provide leverage that is unavailable in observational studies for understanding how specific conceptions of confidence shape analysis and decision making.

A second drawback with survey experiments is that respondents often devote limited attention to survey tasks. In our context, however, limited attention only raises the challenge of obtaining clear results. To the extent that analysts and decision makers were able to engage with abstract elements of uncertainty in low-stakes survey experiments, this makes it less plausible to argue that professionals with greater experience, training, and incentives for effort could not also discriminate among our ideas.

Third, scholars must consider how survey respondents differ from broader populations of interest (Hafner-Burton, Hughes, & Victor, 2013; Hyde, 2015). When studying how our three conceptions of confidence shape decision making, we addressed this problem by recruiting an
elite sample of 223 national security professionals enrolled in mandatory mid-career education at the National War College. The National War College draws active-duty officers from all U.S. military services, along with professionals from civilian national security agencies (32 percent of respondents), and military officers from countries besides the United States (12 percent of respondents). Response rates at advanced military education programs are also high (ours was 88 percent). Our first survey experiment thus comprised a relatively broad cross-section of national security professionals with much less exposure to response bias than most other elite studies.

To study how political analysts engage with our conceptions of confidence, we recruited 1,000 respondents via Amazon Mechanical Turk (AMT). AMT is an online platform where respondents complete surveys in exchange for compensation. Few AMT respondents possess either special interest or special expertise in analyzing national security issues, and we provided no training for assessing uncertainty within the context of our survey. In the context of our

3 We administered this survey between August 9-17, 2016. Mid-career programs like the National War College are required for U.S. military officers gaining promotion to the rank of colonel or commander.

4 On the use of AMT for political science research, see Berinksy, Huber, & Lenz, 2012. We administered this survey experiment on June 30, 2017. Respondents were forty-six percent female and eighty percent white. Forty-seven percent had a college degree.

5 Respondents were compensated $1.40 for finishing a survey that had a median completion time of eleven minutes, a rate of hourly compensation that exceeds the federal minimum standard.

6 On average, respondents reported spending less than one hour per day reading, watching, or discussing political news.
research, however, this should only increase the challenges of confirming that analysts can consistently discriminate among our three conceptions of confidence. All else being equal, we expect that AMT respondents will find it more difficult to disentangle these ideas than would professional analysts with greater training and expertise in assessing uncertainty.

**How national security professionals interpret analytic confidence**

Our first survey experiment employed a vignette-based design, presenting respondents with three scenarios involving national security decisions under uncertainty. These scenarios were a proposed hostage rescue mission with uncertainty about whether the hostages were being held in a particular location; a proposed drone strike with uncertainty about whether the target was a high-ranking terrorist; and a proposal to restrict air travel in response to a reported terrorist plot with uncertainty about whether this threat was real. See supporting information for the full text of each scenario.

Each vignette contained one paragraph describing the decision problem and one paragraph presenting assessments of uncertainty. The second paragraph began with a probability estimate: for example, the chances that the target of the proposed drone strike was a high-ranking terrorist. We randomly assigned one of two values to these probability assessments, which differed by fifteen percentage points. Then we briefly presented our three conceptions of confidence in random order. Each of those confidence assessments took one of two possible values. For example, here is the full text of the hostage rescue scenario, with experimental manipulations in brackets.
The U.S. military is searching for five American citizens held hostage by a rebel group overseas. They receive information suggesting that the hostages are being held in a rural compound. Analysts can tell that the compound is being used by the rebel group, but they have difficulty confirming that the hostages are present. Special forces officers expect that a raid on the compound will meet armed resistance. They do not want to put their soldiers in harm’s way if the hostages are not present, but if they delay action too long, the rebels might move the hostages to a different location.

After careful deliberation, a group of intelligence analysts assesses that there is a [60 / 75] percent chance that the hostages are being held inside this compound. The analysts explain that they have [little / a large amount of] reliable evidence on which to base their judgment; that there is [minimal / significant] disagreement among them about the chances that the hostages are present; and that they [do not believe their assessment would substantially change / believe their assessment could substantially change] if they continue to investigate the compound.

We designed this survey to test three hypotheses. Our principal hypothesis was that each conception of confidence would independently shape the manner in which respondents evaluated national security decisions. This is a demanding hypothesis. Reliability of available evidence, range of reasonable opinion, and responsiveness to new information are abstract ideas. If survey respondents reacted to these ideas simultaneously and independently, even when presented with that information in brief survey vignettes, then this would indicate not only that our three conceptions of confidence provide relevant information in principle, but also that national security professionals intuitively use that information when evaluating decisions under uncertainty.

Our theoretical framework also suggests that respondents should prioritize different conceptions of confidence when evaluating different aspects of a national security decision. For
example, the paper’s first section explained how responsiveness to new information is most important for evaluating the tradeoff between acting immediately and delaying a decision. By contrast, we argued that the reliability of available evidence and the range of reasonable opinion are primarily useful for evaluating a decision’s political and ethical dimensions. Thus our second hypothesis was that respondents would assign responsiveness to new information its greatest weight when evaluating the tradeoff between action and delay, and our third hypothesis was that respondents would give the greatest weight to reliability of available evidence and range of reasonable opinion when evaluating a decision’s political and ethical dimensions. Confirming that respondents prioritize different conceptions of confidence in assessing these different elements of a decision would show that these respondents not only paid attention to this information, but also that they engaged with these ideas in distinctive ways that match theoretical expectations. We summarize our expectations as follows:

**Hypothesis 1.** All three conceptions of confidence will independently influence the way that respondents evaluate proposed decisions under uncertainty.

**Hypothesis 2.** Responsiveness to new information will be most impactful in shaping respondents’ perceptions about the tradeoffs between acting immediately and gathering additional information.

**Hypothesis 3a.** Reliability of available evidence will be most impactful in shaping respondents’ perceptions about the political and ethical dimensions of decisions under uncertainty.

**Hypothesis 3b.** Range of reasonable opinion will be most impactful in shaping respondents’ perceptions about the political and ethical dimensions of decisions under uncertainty.

To test these hypotheses, we asked respondents to indicate their levels of agreement with four statements after reading each scenario: “Decision makers should approve this proposal;”
“Decision makers should gather more information before making a choice;” “It is ethically problematic to take action based on the information available in this scenario;” and “It would be difficult to defend this action to the public.” Supporting information contains full statistical distributions for each measure. We elicited levels of agreement/disagreement with each statement on a seven-point scale.

Results

We analyzed survey data using ordinary least squares regression with fixed effects for each respondent and scenario. (Supporting information contains full results and alternative specifications using ordered logit.) Figure 3 summarizes our principal findings. It shows how our four experimental treatments influenced each of the ways that we asked respondents to evaluate decisions under uncertainty.

[Figure 3]

The results of this survey experiment confirmed Hypothesis 1, as assessments of probability, reliability of available evidence, range of reasonable opinion, and responsiveness to new information all simultaneously influenced the manner in which respondents evaluated proposed courses of action. Each of these four variables was a statistically significant ($p<0.05$) predictor of at least two of the four response measures we elicited. On the central question of the extent to which respondents supported taking action in each case, assessments of probability, range of reasonable opinion, and responsiveness to new information each exerted effects that were statistically significant at the $p<0.05$ level, while reliability of available evidence narrowly missed that threshold ($p=0.06$).
Our experimental results also confirmed Hypothesis 2, showing that respondents gave responsiveness to new information its greatest weight when evaluating tradeoffs between acting immediately and gathering additional information. The substantive impact of responsiveness to new information was twice as large for predicting willingness to delay action ($\beta=-0.59$) than for any of the other three measures we elicited from respondents: the next highest is support for taking action ($\beta=0.27$).

Our experiment produced mixed support for Hypotheses 3a and 3b. As anticipated, we found that the reliability of available evidence had its greatest impact for shaping respondents’ views about the extent to which a national security decision would be politically controversial. And our results also demonstrate that respondents believed that range of reasonable opinion was important for evaluating the political dimensions of their decisions ($\beta=-0.29$, $p<0.01$). However, contrary to our expectations, we found that range of reasonable opinion had its greatest impact in shaping respondents’ willingness to gather additional information ($\beta=-0.46$, $p<0.001$).

To explore this unexpected finding further, we analyzed how respondents combined information about range of reasonable opinion and responsiveness to new information when evaluating tradeoffs between acting immediately and gathering additional information. Table 1 shows that respondents were most willing to delay action when analysts reported that their findings could change substantially in light of new information. In this context, range of reasonable opinion had little statistical impact on respondents’ views ($\beta=0.12$, $p=0.53$). By contrast, range of reasonable opinion explained substantial variation in respondents’ willingness to delay action when analysts reported that their findings were unlikely to change in response to new information ($\beta=0.70$, $p<0.001$).
These patterns suggest that respondents applied a two-step decision rule when considering tradeoffs between the benefits of gathering additional information versus the potential costs of delay. The primary conception of confidence that respondents examined when considering this tradeoff in our experiment was the extent to which delaying action would produce informational gains. However, even when respondents were skeptical about the informational value of delay, they were still hesitant to act as long as there were unresolved differences over how to assess uncertainty. This pattern of behavior indicates that, even in brief scenarios and without explicit instructions, respondents engaged with multiple conceptions of confidence consistently and distinctively. As we discuss in the paper’s conclusion, one potentially productive avenue for future research on this subject would be to conduct further experiments that examine such thought processes in greater detail.

The results of this survey experiment demonstrate that our three conceptions of confidence capture independent elements of high-stakes decision making, elements that matter not merely in principle but also in practice. Though distinctions among our three conceptions of confidence are abstract, and though we gave respondents no special instructions about how to interpret this information, we found that these attributes influenced the way that national security professionals intuitively evaluated decisions under uncertainty. Not only did our respondents react to these cues, but they also prioritized different elements of analytic confidence when assessing specific elements of decision making in ways that either matched our theoretical expectations or suggested otherwise sensible approaches to managing uncertainty.
How inexperienced analysts assess analytic confidence

Even if our three conceptions of confidence independently shape the way that national security officials evaluate high-stakes decisions, these ideas would still offer little practical value if analysts could not reliably distinguish among them. To investigate how individuals can cope with this challenge, we administered a second survey experiment to 1,000 respondents via Amazon Mechanical Turk (AMT). We asked each respondent to answer ten questions, randomly chosen from a broader list of twenty. Eight questions pertained to U.S. domestic politics, such as “What are the chances that Hillary Clinton will win the 2016 U.S. presidential election?” Eight questions pertained to foreign policy issues, such as “What are the chances that Britain will formally exit the European Union by the end of 2016?” The remaining questions dealt with issues that involved both foreign and domestic politics, such as “What are the chances that the United States will accept more than 7,500 Syrian refugees by the end of 2016?” Figure 4 summarizes these questions: see supporting information for complete wordings.

For each question, we asked respondents to provide a probability estimate. Then we asked respondents to indicate the extent to which they agreed with each of the following statements: “I could defend this estimate with a substantial amount of reliable information;” “Reasonable people could give substantially different answers to this question;” and “My answer could change substantially if I studied this subject further.” We elicited agreement with these statements on five-point scales (strongly disagree, disagree, neutral, agree, strongly agree), as this is the level of precision with which intelligence analysts generally assign confidence levels (low, low-to-moderate, moderate, moderate-to-high, high). After gathering these data, we ordered each scale so that larger values indicated higher levels of confidence.
Identifying independent variation

As our supporting information describes in greater detail, respondents assigned the same value to all three conceptions of confidence on just seventeen percent of observations. For twenty-five percent of observations, the confidence levels that respondents attached to their forecasts differed by at least three points out of five across measures. No two conceptions of confidence exhibited a correlation coefficient greater than 0.53.7 These descriptive statistics show that there was substantial independent variation in how respondents assessed our three conceptions of confidence.

Nevertheless, this independent variation could reflect arbitrary responses to survey questions as opposed to systematic distinctions among abstract ideas. If the relationship between our three measures tended to be stable on average, and if the remaining variation was simply random noise, then there would be little practical value in asking analysts to distinguish among these concepts. The key empirical issue is thus whether respondents’ confidence levels varied across measures in a manner that was both independent and consistent.

To evaluate this claim, we analyzed variation in how respondents assessed each conception of confidence while controlling for the following factors: (i) the degree of certainty that respondents assigned to each forecast, defined as the absolute value of the difference between each

---

7 The correlation between reliability of available evidence and range of reasonable opinion was 0.27; the correlation between reliability of available evidence and responsiveness to new information was 0.34; and the correlation between range of reasonable opinion and responsiveness to new information was 0.53.
probability estimate and fifty percent; (ii) how respondents assessed the other two kinds of confidence when making each forecast; and (iii) demographic variables, including indicators for whether respondents were female, white, and/or college-educated, respondents’ age in years, a seven-step measure of respondent’s political ideology ranging from very conservative to very liberal, and respondents’ political engagement as measured by the number of hours per week that they reported watching, reading, or discussing political news.

After controlling for these measures, we added a dummy variable for each of our survey questions, withholding the twentieth question as a baseline. The coefficients on these dummy variables indicate the extent to which respondents’ assessments of confidence varied across questions, even after controlling for probability estimates, individual attributes, and the way that respondents assessed other conceptions of confidence. If these question-level fixed effects prove to be statistically significant, this would confirm that eliciting conceptions of confidence provided information that was not contained in other aspects of how they assessed uncertainty.

Results

Figure 4 presents results, estimated using ordinary least squares regression with respondent fixed effects and standard errors clustered by respondent. (See supporting information for full analysis and additional specifications using ordered logit.) Each column presents a statistical model predicting a different conception of confidence. For only one question did we find no statistically significant difference between the way that respondents assessed each conception of confidence and what we would have predicted given other covariates. Nearly two-thirds of the question-level fixed effects examined in Figure 4 (37 out of 57, or 65 percent) are statistically
significant at the $p<0.05$ level, and two-fifths (23 out of 57, or 40 percent) are statistically significant at the $p<0.001$ level.

[Figure 4]

To illustrate our findings, Figure 4 shows that when analysts estimated the chances that Al Qaeda’s leader would be killed or captured by the end of the year (question 19), they accepted that this judgment was based on an unusually limited amount of reliable information ($\beta=-0.58$, $p<0.001$), even after controlling for the way that those respondents evaluated probability and the other two kinds of confidence in describing their beliefs. When respondents estimated the chances that Barack Obama’s approval rating would be above fifty percent by the end of 2016 (question 3), they described this judgment as being unusually open to reasonable disagreement ($\beta=-0.42$, $p<0.001$). When respondents estimated the chances that the U.S. Congress would pass gun control legislation by the end of 2016 (question 7), they believed that their views were particularly unlikely to change in response to new information ($\beta=0.29$, $p<0.001$). For six of twenty questions, we observed independent and statistically significant departure from the mean for all three conceptions of confidence at once. For example, when respondents estimated the chances that Hillary Clinton would be elected president (question 1), they generally reported that their beliefs were based on unusually large amounts of reliable evidence, and that those beliefs were unusually unlikely to change in response to new information, but that reasonable people were also unusually likely to disagree when offering their views on this issue.

These results demonstrate that our respondents consistently distinguished among the ways they assessed each conception of confidence when making political forecasts. Though our survey respondents had no particular experience analyzing politics, and though we provided them with
no special instructions about how to disentangle abstract elements of assessing uncertainty, we found that they could still intuitively discriminate among these concepts. Thus while our first survey experiment showed that these distinctions matter to the ways that national security professionals evaluate decisions under uncertainty, our second survey suggests that even inexperienced assessors can draw these distinctions in practice.

Summary and directions for further research

In 2002, the United States Intelligence Community released a National Intelligence Estimate, judging with “high confidence” that Iraq was “continuing, and in some areas expanding, its chemical, biological, nuclear and missile programs.” What did it mean to make this judgment with high confidence? One interpretation is that analysts thought it was highly likely that Iraq was pursuing weapons of mass destruction (WMDs). Perhaps this statement was ninety percent likely to be true: a “slam dunk” as this statement was described in public debate. Or maybe analysts thought that the chances of Iraq pursuing weapons of mass destruction were more like sixty percent, and they used the term “high confidence” to indicate that most analysts believed this judgment was likely to be correct. While the difference between these interpretations could have had markedly different implications for debates about invading Iraq, there was little discussion at the time, either in government or in the public sphere, about what exactly it meant to assess uncertainty with “high confidence” (Jervis, 2010, ch. 3).

Over the last decade, intelligence agencies in the United States and in several other countries have undertaken substantial efforts to reduce confusion when communicating uncertainty (Fingar, 2011; Ho et al. 2015; Wheaton, 2012). Yet scholars and practitioners still lack
conceptual foundations for describing what analytic confidence means in principle, and there remain open empirical questions about the extent to which analysts and decision makers can engage with these ideas in practice. To address these issues, we argued for disaggregating “analytic confidence” into three distinct attributes: reliability of available evidence, range of reasonable opinion, and responsiveness to new information. We showed that these attributes can vary independently and explained why they hold different implications for high-stakes decision making. Our survey experiments demonstrated that these ideas shape the way that national security professionals evaluate decisions under uncertainty, and that even inexperienced analysts can consistently discriminate among these abstract ideas. Though our analysis focused primarily on the challenges of assessing uncertainty in national security, we explained how similar issues surround virtually any other area of political decision making.

Further research can extend our analysis in at least three ways. One avenue for additional research would explore how different conceptions of confidence covary across issues. It is particularly important to determine the conditions under which our three conceptions of confidence are most likely to diverge. These are the cases where separating conceptions of confidence would likely provide the most practical value for shaping decisions under uncertainty.

A second extension of this research would be to use within-case process tracing to explore how decision makers respond to different conceptions of confidence when making high-stakes choices. Since survey experiments cannot capture the dense mixture of cues and pressures that real high-stakes decision entail, we have limited basis for making clear predictions about exactly where and when these conceptions of confidence have their greatest impact. Further work on this subject – perhaps delving into the cases we described in the paper’s first section like analyses of
Iraq’s weapons of mass destruction programs or the search for Osama bin Laden – would hone scholars’ ability to describe how different components of uncertainty shape major decisions.

A third potential extension of this research program would explore the psychological mechanisms by which decision makers interpret and combine our three conceptions of confidence. Though many scholars question the extent to which individuals can grapple with the nuances of assessing uncertainty, our survey experiments showed how both elite and non-elite respondents could consistently separate probability from confidence, and that they could reliably disentangle confidence into three distinct components. Our first survey experiment further suggested that decision makers interpreted responsiveness to new information and range of reasonable opinion using a sensible two-step process. Altogether, these findings offer encouraging implications regarding individuals’ capabilities to engage with abstract elements of uncertainty, inviting further experimental investigation of what these thought processes entail.
References


Expressing Analytic Certainty

Certainty is based on two distinct components: likelihood and confidence. This chart is designed to assist readers in understanding the language we use in our products.

**Likelihood**

<table>
<thead>
<tr>
<th>Will Not</th>
<th>Unlikely</th>
<th>Unlikely Improbable</th>
<th>Unlikely Highly Improbable</th>
<th>Very Doubtful</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Do not foresee</td>
<td>Do not anticipate</td>
<td>Do not expect</td>
<td>Does not appear doubtful</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Confidence**

<table>
<thead>
<tr>
<th>Decreasing Likelihood</th>
<th>Confidence</th>
<th>Increasing Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Well-corroborated information from proven sources, low potential for deception, nonhistorical assumptions and/or gaps, or undisputed reasoning.</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>Partially corroborated information from good sources, moderate potential for deception, potentially critical assumptions used to fill gaps, or a rate of inference.</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Uncorroborated information from good or marginal sources, high potential for deception, key assumptions used to fill critical gaps, or mostly weak inferences.</td>
<td></td>
</tr>
</tbody>
</table>

**Estimative Language**

Estimative language consists of two elements: judgments about the likelihood of developments or events occurring, and levels of confidence in the source and analytic reasoning supporting the judgments. Judgments are not intended to imply that we have proof that shows something to be a fact. Assessments are based on collected information, which is often incomplete or fragmentary, as well as logic, argumentation, and precedent.

**Judgments of Likelihood.** The chart below approximates how judgments of likelihood correlate with percentages. Unless otherwise stated, the Intelligence Community’s judgments are not derived via statistical analysis. Phrases such as “we judge” and “we assess”—and terms such as “probable” and “likely”—convey analytical assessments.

<table>
<thead>
<tr>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

**Confidence in the Sources Supporting Judgments.** Confidence levels provide assessments of the quality and quantity of the source information that supports judgments. Consequently, we ascribe high, moderate, or low levels of confidence to assessments.

- **High confidence** generally indicates that judgments are based on high-quality information from multiple sources. High confidence in a judgment does not imply that the assessment is a fact or a certainty; such judgments might be wrong.
- **Moderate confidence** generally means that the information is credibly sourced and plausible but not of sufficient quality or corroborated sufficiently to warrant a higher level of confidence.
- **Low confidence** generally means that the information’s credibility and/or plausibility is uncertain, that the information is too fragmented or poorly corroborated to make solid analytic inferences, or that reliability of the sources is questionable.

2015 Defense Intelligence Agency Guidelines

2017 National Intelligence Council Guidelines

Figure 1. Instructions for expressing uncertainty in intelligence
Friedman and Zeckhauser, “Analytic Confidence”

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Evidence (type, amount, quality, consistency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High agreement</td>
<td>Limited evidence</td>
</tr>
<tr>
<td>Medium agreement</td>
<td>Limited evidence</td>
</tr>
<tr>
<td>Low agreement</td>
<td>Limited evidence</td>
</tr>
<tr>
<td>High agreement</td>
<td>Medium evidence</td>
</tr>
<tr>
<td>Medium agreement</td>
<td>Medium evidence</td>
</tr>
<tr>
<td>Low agreement</td>
<td>Medium evidence</td>
</tr>
<tr>
<td>High agreement</td>
<td>Robust evidence</td>
</tr>
<tr>
<td>Medium agreement</td>
<td>Robust evidence</td>
</tr>
<tr>
<td>Low agreement</td>
<td>Robust evidence</td>
</tr>
</tbody>
</table>

Figure 2. IPCC guidelines for assessing confidence
Figure 3 presents estimated treatment effects from OLS regressions characterizing respondents’ reactions to national security decisions. Dependent variables measured on 7-point scales. Supporting information contains full analysis.

**Figure 3. Responses to national security scenarios**
Figure 4 shows mean (standard error) estimates for how respondents' assessments of each conception of confidence deviated from the norm on each question, controlling for respondents’ assessments of probability, other conceptions of confidence, and demographic attributes. We withheld Question 20 as a baseline, which asked about the chances that more than 10 U.S. soldiers would be killed in Iraq in 2016. ^: designates that the question also specified the time frame was by/at the end of 2016.

**Figure 4. Exploring independent variation in conceptions of confidence across questions**
<table>
<thead>
<tr>
<th></th>
<th>High responsiveness</th>
<th>Low responsiveness</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substantial disagreement</td>
<td>5.32 (1.72)</td>
<td>5.01 (1.77)</td>
<td>0.31 (p=0.10)</td>
</tr>
<tr>
<td>Little disagreement</td>
<td>5.19 (1.83)</td>
<td>4.31 (1.84)</td>
<td>0.88 (p&lt;0.001)</td>
</tr>
<tr>
<td>Difference</td>
<td>0.12 (p=0.53)</td>
<td>0.70 (p&lt;0.001)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 shows mean support for delaying action (standard deviations in parentheses) based on assessments of range of reasonable opinion and responsiveness to new information; p-values reflect comparisons of means in two-way t-tests.

**Table 1. How respondents reacted to range of reasonable opinion vs. responsiveness to new information**