CONSTRUCTING AN ELICITATION ON THE RISKS OF WEAPONS OF MASS DESTRUCTION

Lessons from Analyzing the Lugar Survey

National Security Report

Jane Booker | Lori Baxter | James Scouras

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

CONSTRUCTING AN ELICITATION ON THE RISKS OF WEAPONS OF MASS DESTRUCTION

Lessons from Analyzing the Lugar Survey

Jane Booker

Lori Baxter

James Scouras

APL

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

COMMS-23-02417

Contents

Figures

Tables

Summary

The perceived risk of weapons of mass destruction (WMD) has been a critical driver of US national security policy since the Second World War when fear of Germany developing atomic weapons galvanized support for the Manhattan Project. While this assessment applies generally to biological, chemical, and radiological attacks, it is most clearly evident in the case of nuclear weapons, where significant policy initiatives have been motivated by concerns that deterrence might fail. Examples include the "hotline" agreement and the ensuing edifice of nuclear arms control agreements, the Strategic Defense Initiative, and the Global Zero initiative.[1] Yet, the underlying risk perceptions that drive such initiatives have been largely intuitive and, thus, of uncertain validity. Attempts to quantify risks more accurately have generally relied on expert judgments, but this approach has proven fraught with perils.

Not much has changed since the Manhattan Project and initial fears associated with Soviet demonstration of a nascent nuclear capability in 1949. While we now know much more about consequences of WMD attacks, significant uncertainties remain.[2] Most important, assessing the likelihoods of such attacks remains an elusive endeavor because of the sparsity of data, which results in heavy reliance on the information, experience, and knowledge of experts in diverse fields. In this paper, we critically evaluate one such assessment with the hope that our work will help elevate the quality of similar future efforts.

In 2005, Senator Richard Lugar polled eighty-five experts to quantify their WMD risk perceptions and identify points of convergence and divergence, with the overarching goal of drawing increased attention to the need for greater nonproliferation efforts. The results of this survey are presented in *The Lugar Survey on Proliferation Threats and Responses*, and the contents of that report are the focus of our analysis efforts. We also examine nearly two decades of citations of this report from 2005 to mid-2023. An online appendix summarizes our literature search.

Senator Lugar was clear in cautioning that the survey was a political effort, rather than a "scientific" survey. And it has been successful in its goal of encouraging dialogue on proliferation threats. We find, however, that most documents citing the Lugar survey have ignored Senator Lugar's caution by taking its results at face value—in other words, without appropriate caveats—thereby lending greater credibility to it than is warranted. Our examination of the Lugar survey explains the basis for Senator Lugar's caution, focusing on survey methodology and implementation, as well as the analysis and presentation of results. We offer numerous suggestions for any future survey that aspires to utilize best elicitation and analysis practices.

[1] Scouras, *On Assessing the Risk*, 6–8.

[2] Frankel, Scouras, and Ullrich, *Uncertain Consequences*.

Introduction

Deterring attacks using weapons of mass destruction (WMD)—chemical, biological, radiological, and nuclear (CBRN)—and preventing proliferation of such weapons to additional state and non-state entities are pillars of US national security strategy. To help develop effective and balanced policies in these critical areas, it is necessary to realistically assess the threats—or risks—of various future adverse developments, as well as the efficacy of current and prospective actions we might take to reduce these risks.

Unfortunately, accurate CBRN risk assessments are extremely challenging as they involve complex interactions of multiple human and institutional actors with differing personalities, objectives, and constraints. Moreover, adverse developments often turn on the vagaries of fate, as so many past incidents and close calls have amply demonstrated. Data are sparse, analytical methodologies are coarse, and the future is never a straightforward extrapolation of the past. In such circumstances, we often turn to elicitation of experts to provide information and knowledge.

The Lugar Survey on Proliferation Threats and Responses[1] (hereinafter referred to as the Lugar survey report) does just that. A sample survey on CBRN risks, directed by the late Senator Richard Lugar and conducted in 2004–2005, poses twenty-one questions to eighty-five responding experts. Its results have been used by Senator Lugar and others in policy formulation and have been widely cited in academic and nongovernmental organization (NGO) literature.

Never intended to be a “scientific” survey, the Lugar survey was nevertheless a pioneering effort that provides useful lessons for similar future surveys that do aspire to be analytically rigorous. Thus, the purpose of this paper is to draw such lessons to inform future expert surveys of WMD risks. Our report is organized into the following sections, each of which can be read independently of the others as desired:

(1) Background on the Lugar Survey provides a short description of the Lugar survey motivations, design, and survey instrument.

(2) Analysis of the Lugar Survey Results contains statistical analyses of the Lugar survey results and conclusions and offers suggestions for additional statistically based presentations.

(3) Uses and Abuses of the Lugar Survey documents the survey’s usage in the literature. This section is expanded in the online appendix to this paper.[2]

(4) Formal Elicitation of Expert Knowledge Topics[3] offers guidance for planning and designing a more rigorous study with a defensible foundation for a knowledge base.

(5) Final Thoughts provides a summary and a discussion on the path forward.

Background on the Lugar Survey

Nonproliferation

The purpose of the Lugar survey was to raise awareness of nonproliferation issues. As indicated by the title of the first major section in the Lugar survey report, Building on Existing Non-Proliferation Efforts, it was one of several efforts to do this, including speeches, op-eds, and legislation. The major legislative program focused on the Nunn-Lugar Act, coauthored by Senators Sam Nunn (D-GA) and Richard Lugar (R-IN).[4] Policy initiatives included the US nuclear test moratorium implemented by a

[1] Lugar, *Lugar Survey.*

[2] Published online at https://www.jhuapl.edu/sites/default/files/2024-05/LugarSurveyLiteratureAnalysis.pdf.

[3] Meyer and Booker, *Eliciting and Analyzing Expert Judgment.*

[4] This legislation is also known as the Soviet Nuclear Threat Reduction Act of 1991.

1992 executive order from President George H. W. Bush that remains in effect today. Test ban treaties, both approved and proposed, and nuclear weapon stockpile reductions were also key components of US and international discussion regarding nonproliferation.

The Lugar survey report clearly cautions that it is neither a scientific survey nor a research endeavor: "I would underscore that this study is not meant to be a scientific poll of the entire national security community."[5] Rather, as expressed by Dan Diller, who worked closely with Senator Lugar on the survey and is currently director of policy at the Lugar Center, it was intended to be a "hearing" of experts (identified as "elites"[6]) known to and selected by Senator Lugar and his staff.[7] A total of one hundred thirty-two such experts were identified. After the survey team contacted those who did not initially respond, eighty-five experts participated. Senator Lugar himself called some, encouraging them to respond.

Most experts responded to each question, even though they were given the option to skip questions. For example, question 2 had the fewest responses, with seventy-seven, and question 6 had the most, with all eighty-five. This raises the issue of depth versus breadth of expertise, which we address later.

For many of the questions (4–5 and 9–14), experts were asked to provide probabilities/percentages for WMD attacks for future time periods of five and ten years. Since 2005, those five- and ten-year timelines have long expired; therefore, the event as specified by the questions either did not occur or did occur. For example, question 6 asks for the probability of a nuclear explosion in an attack somewhere in the world in the next ten years. This event has not occurred. In fact, as of this writing, none of the prospective WMD attacks have occurred since 2005. These new (since the survey) data provide an opportunity to assess the predictive capability of the experts.

The Lugar survey represents a pioneering, yet flawed, effort to provide a snapshot of knowledge from a sample of nonproliferation experts in 2005. As such, it offers an opportunity to draw lessons that could guide a future study that is better grounded in state-of-the-art elicitation and analysis methods.

Lugar Survey Questions

For reference throughout this paper, the Lugar survey questions follow. Charts have been remade only for clarity of presentation. To summarize responses, we have included one statistical measure of dispersion (range), as well as the two statistical measures of central tendency (average and median) provided in the original report. The number of respondents is also listed. These are the measures reported or directly observable in Lugar's analysis of results.

Our reproduction of the Lugar survey histograms may be imprecise due to occasional difficulty in reading the original charts. We believe the error due to this effect is no more than ±1 response.

[5] Lugar, *Lugar Survey*, 4.

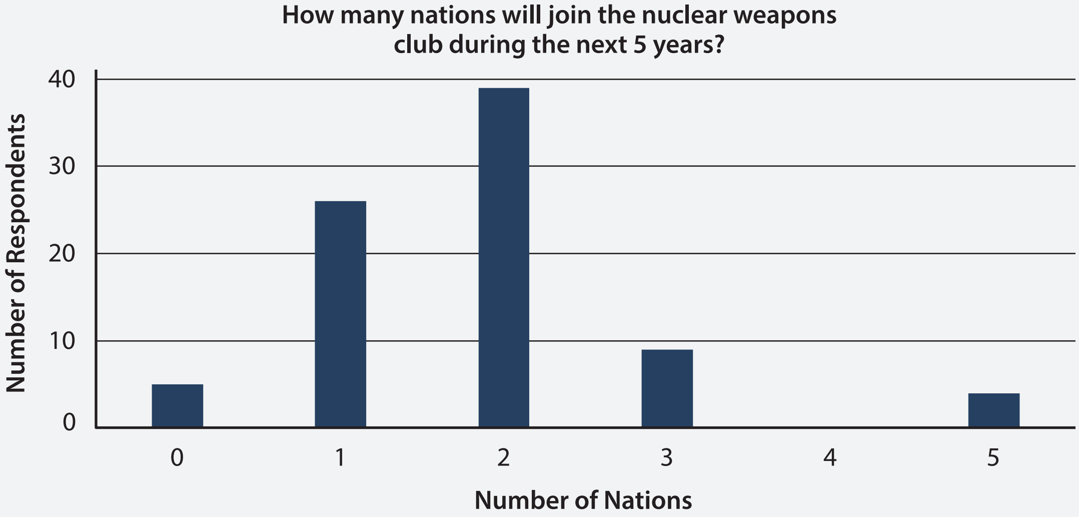
[6] Conversation with Dan Diller, director of policy at the Lugar Center, on August 11, 2022. Diller did not define "elites," nor does the Lugar survey report supply a description of their one hundred thirty-two "non-proliferation and national security experts" (p. 4), but we take this term to refer to average WMD experts that are personally known to Senator Lugar and staff and particularly influential in their eyes.

[7] Conversation with Dan Diller, director of policy at the Lugar Center, on August 11, 2022.

On Nuclear Proliferation

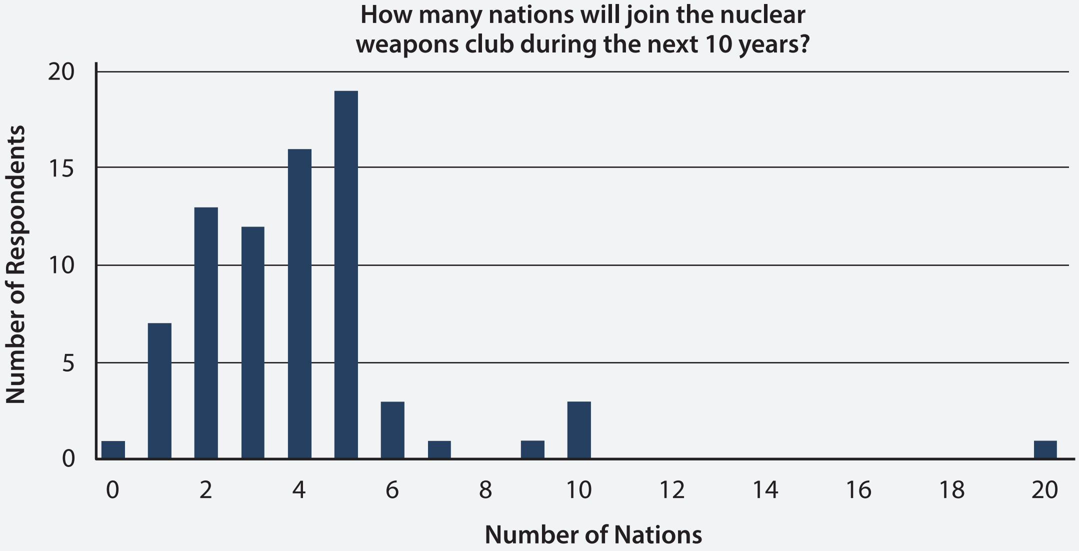
(1) In your estimate, how many nations that do not currently possess a working nuclear weapon will be added to the nuclear weapons club during the next 5 years?

Responses: 83 Range: 0–5 Median: 2 Average: 1.8



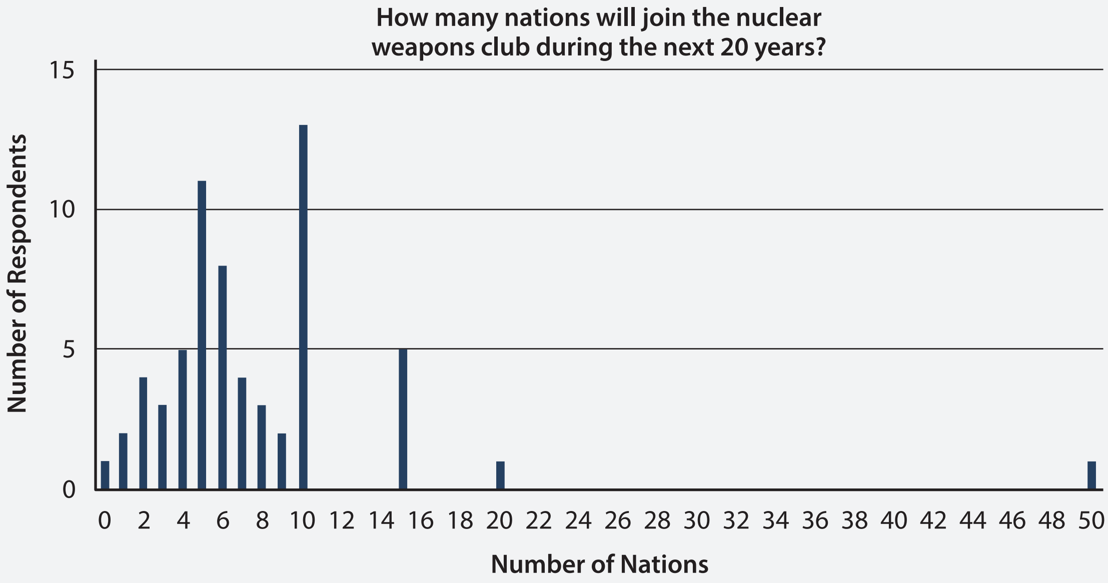
(2) In your estimate, how many nations that do not currently possess a working nuclear weapon will be added to the nuclear weapons club during the next 10 years?

Responses: 77 Range: 0–20 Median: 4 Average: 4



(3) In your estimate, how many nations that do not currently possess a working nuclear weapon will be added to the nuclear weapons club during the next 20 years?

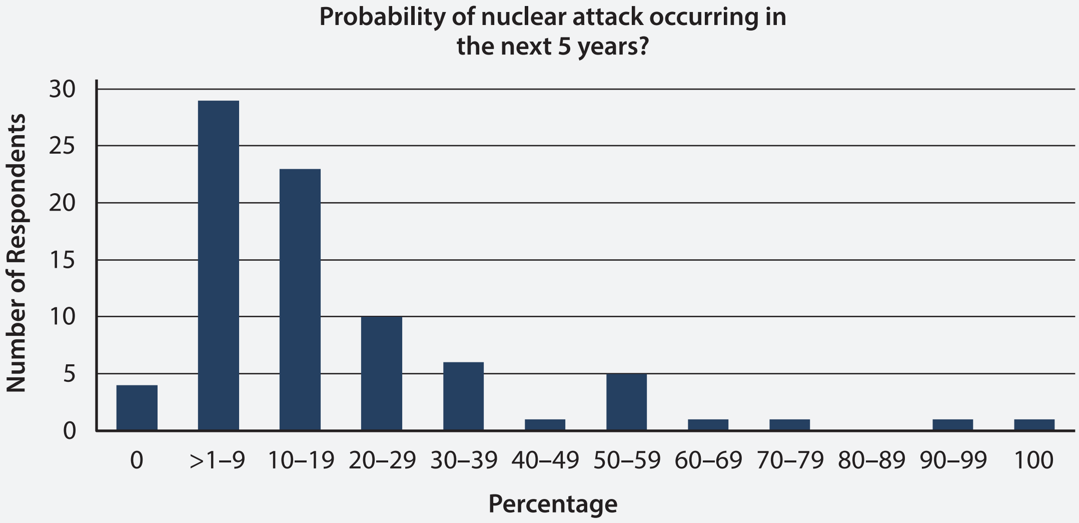
Responses: 63 Range: 0–50 Median: 6 Average: 7.5



On Nuclear Risks

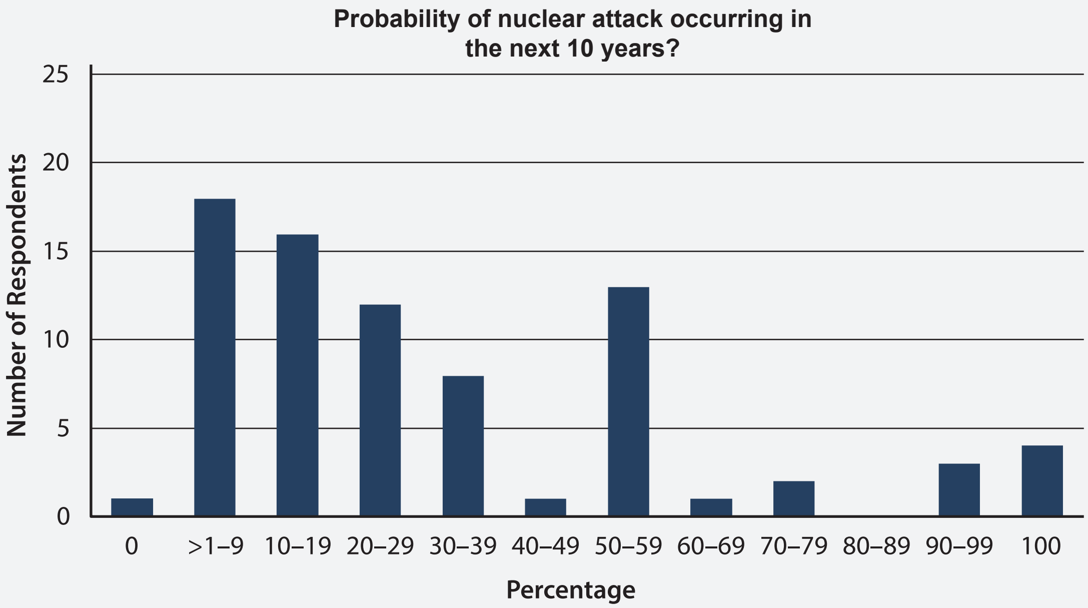
(4) In your opinion, what is the probability (expressed as a percentage) of an attack involving a nuclear explosion occurring somewhere in the world in the next 5 years?

Responses: 82 Range: 0–100 Median: 10% Average: 16.4%



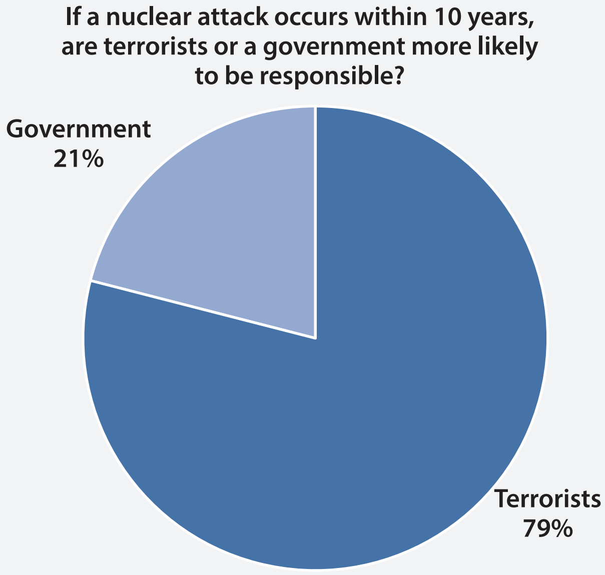
(5) In your opinion, what is the probability (expressed as a percentage) of an attack involving a nuclear explosion occurring somewhere in the world in the next 10 years?

Responses: 79 Range: 0–100 Median: 20% Average: 29.2%



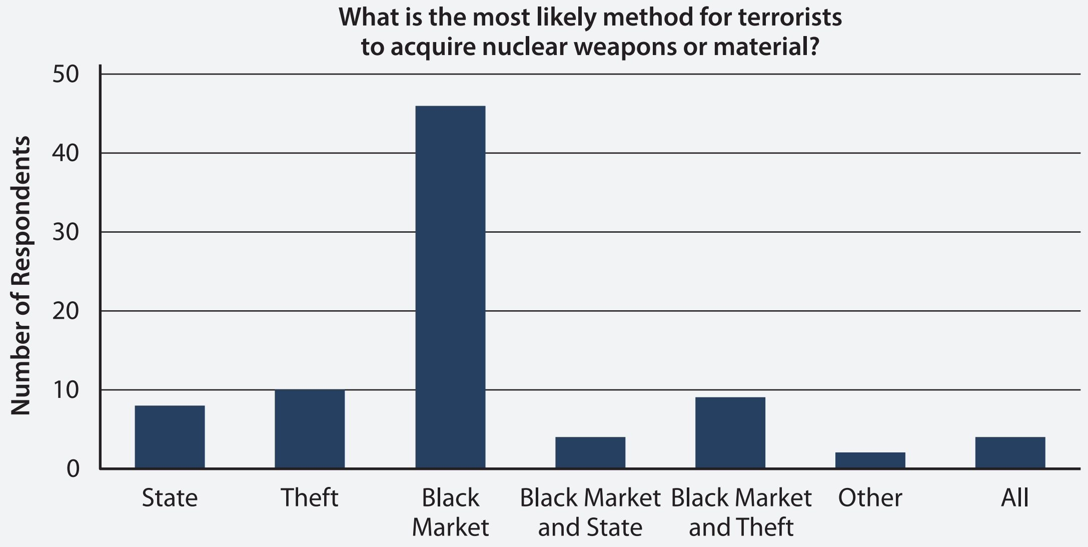
(6) In your opinion, if a nuclear attack occurs during the next 10 years, is it more likely to be carried out by terrorists or by a government?

Responses: 85



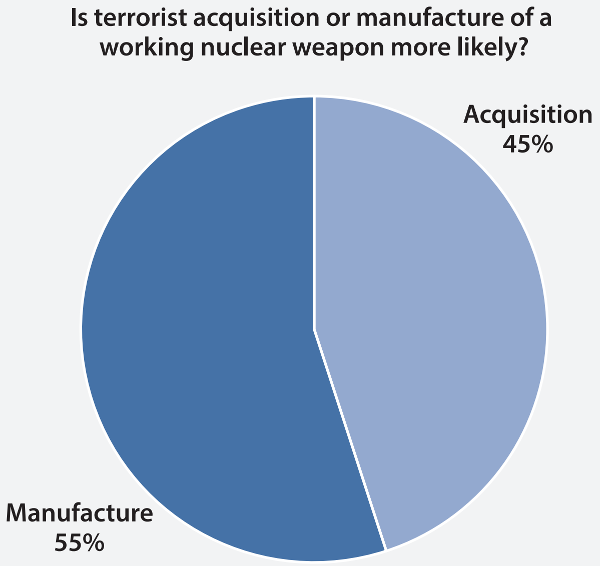
(7) What is the most likely method through which terrorists would acquire nuclear weapons or weapons grade nuclear material: a) theft; b) black market purchase; c) transfer or sale from a nuclear weapons state; d) other?

Responses: 83 Range: N/A Median: N/A Average: N/A



(8) In your opinion, which proliferation scenario is more likely: terrorist acquisition of a working nuclear weapon or terrorist manufacture of a nuclear weapon after obtaining weapons grade nuclear material?

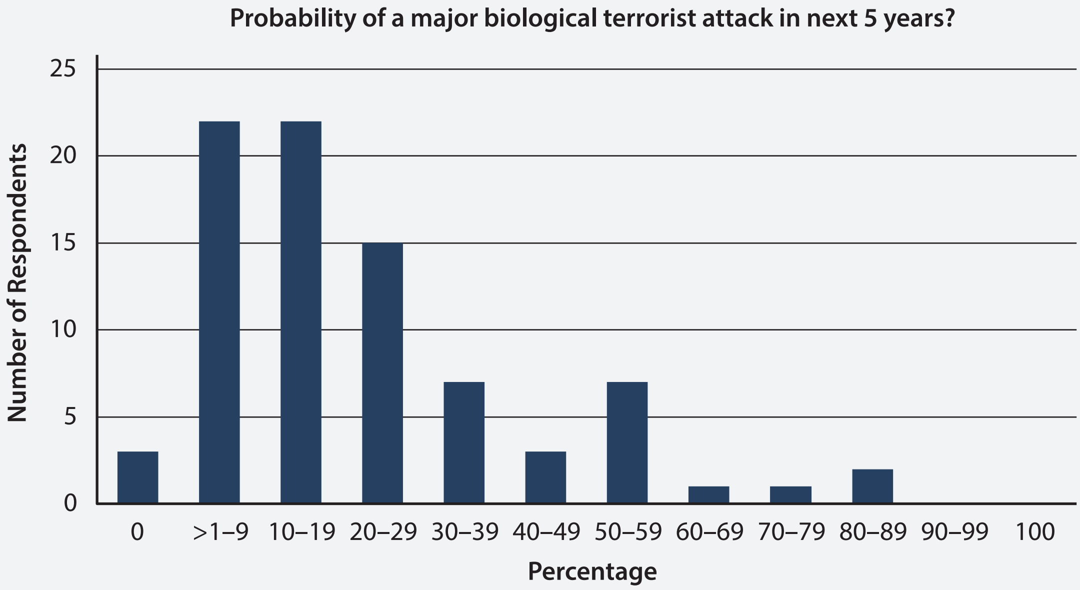
Responses: 82



On Biological Risks

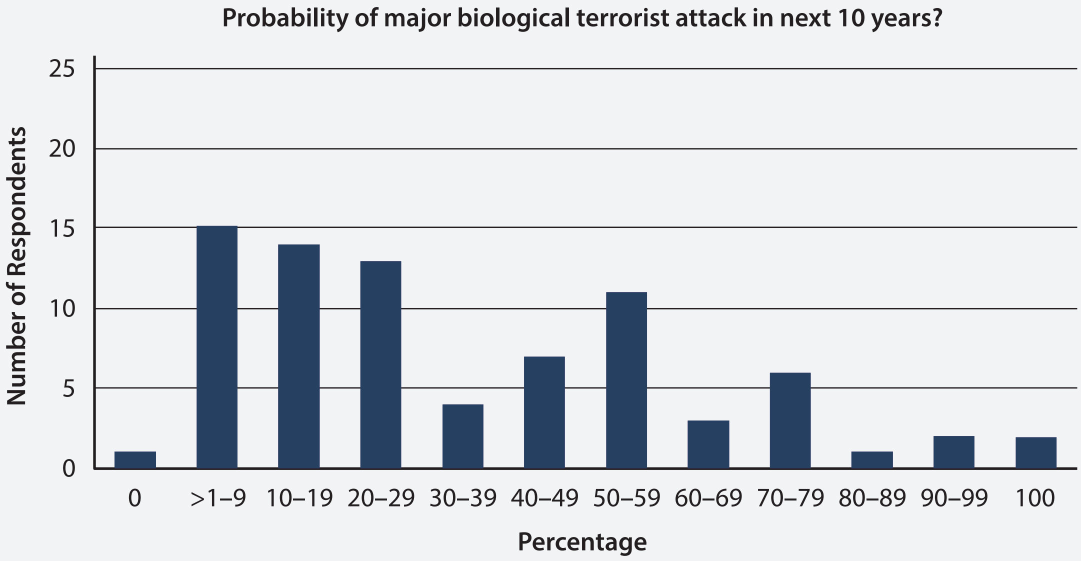
(9) In your opinion, what is the probability (expressed as a percentage) of a major biological terrorist attack that inflicts numerous fatalities in the next 5 years?

Responses: 83 Range: 0–89% Median: 10% Average: 19.7%



(10) In your opinion, what is the probability (expressed as a percentage) of a major biological terrorist attack that inflicts numerous fatalities in the next 10 years?

Responses: 79 Range: 0–100% Median: 20% Average: 32.6%

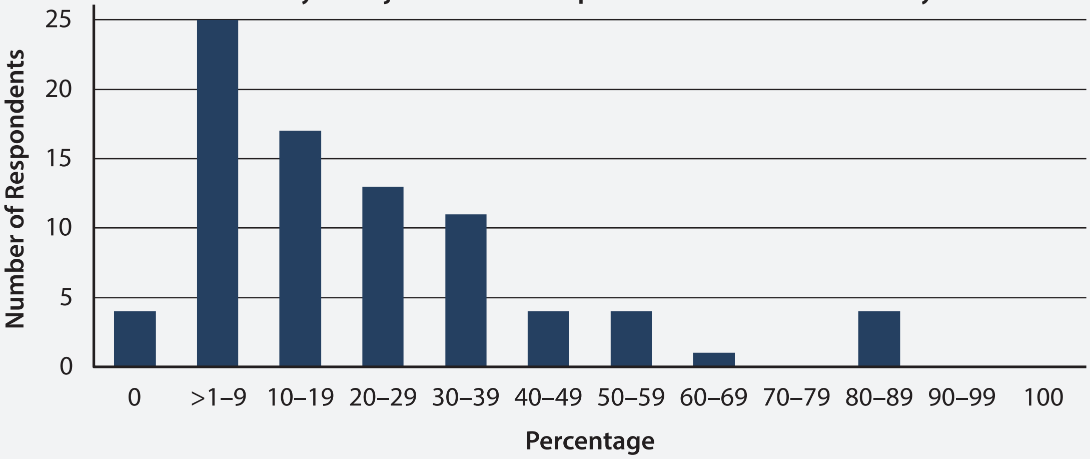


On Chemical Risks

(11) In your opinion, what is the probability of a major chemical weapons terrorist attack that inflicts numerous fatalities in the next 5 years?

Responses: 83 Range: 0–89% Median: 15% Average: 20.1%

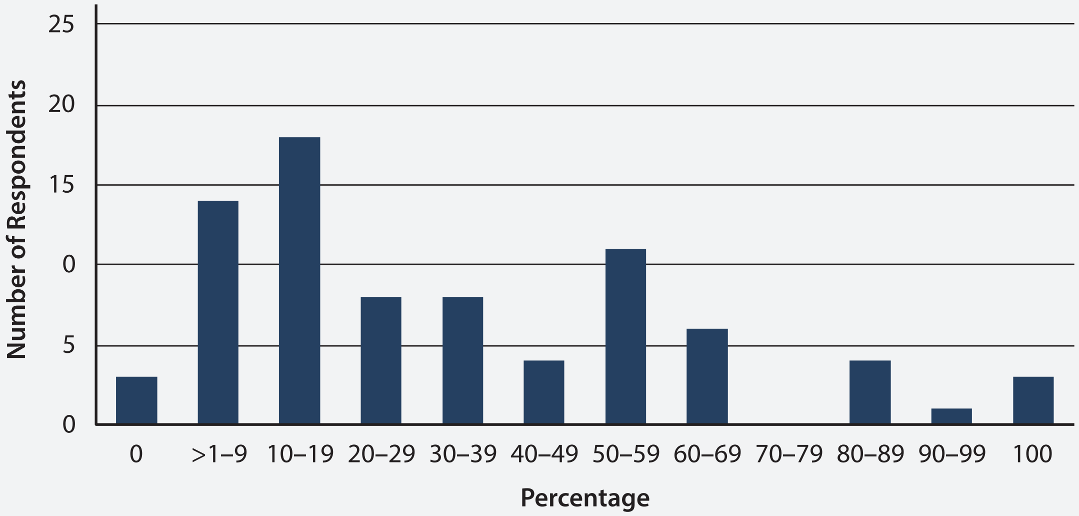
Probability of major chemical weapons terrorist attack in next 5 years?



(12) In your opinion, what is the probability of a major chemical weapons terrorist attack that inflicts numerous fatalities in the next 10 years?

Responses: 80 Range: 0–100% Median: 15% Average: 30.5%

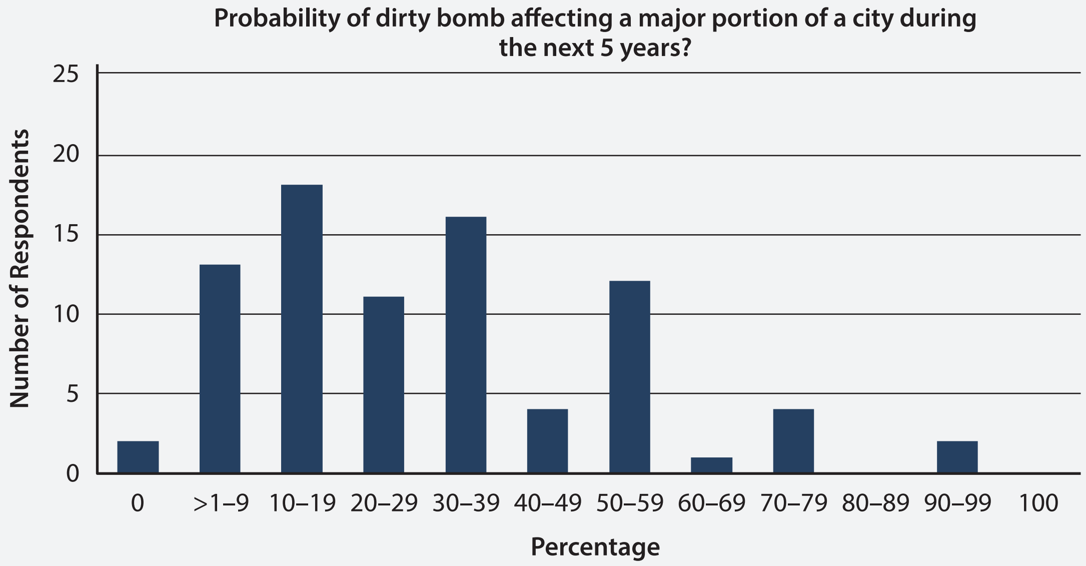
Probability of major chemical weapons terrorist attack in next 10 years?



On Radiological Risks

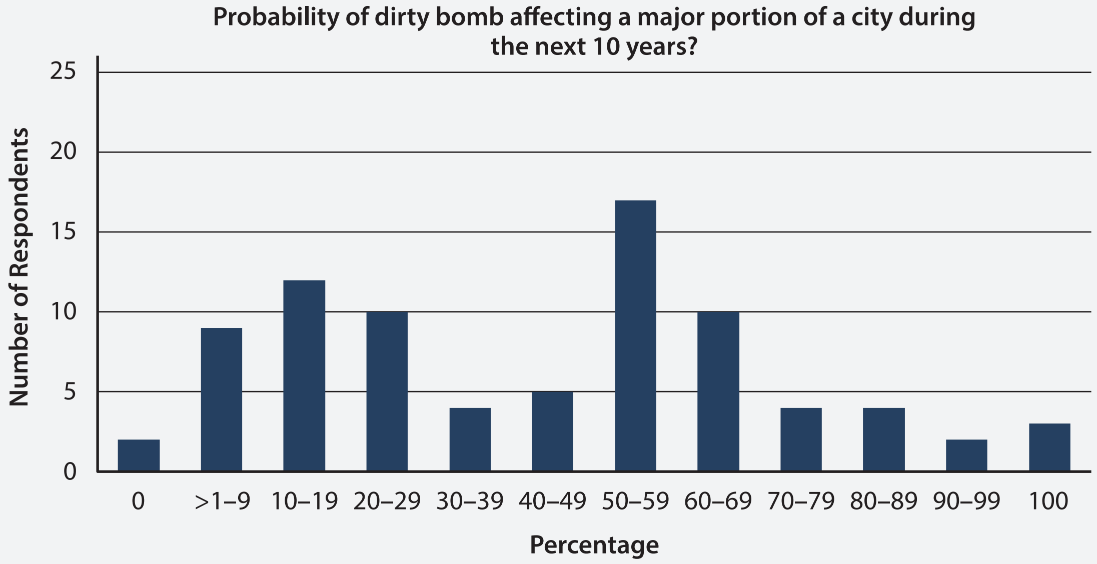
(13) In your opinion, what is the probability of a terrorist attack using a radiological dispersal device (dirty bomb) that affects a major portion of a city during the next 5 years?

Responses: 83 Range: 0–99% Median: 25% Average: 27.1%



(14) In your opinion, what is the probability of a terrorist attack using a radiological dispersal device (dirty bomb) that affects a major portion of a city during the next 10 years?

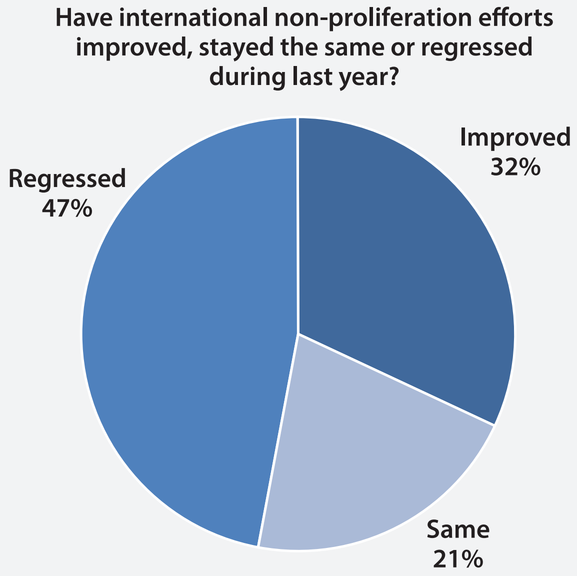
Responses: 82 Range: 0–100% Median: 40% Average: 39.8%



On Nonproliferation Efforts

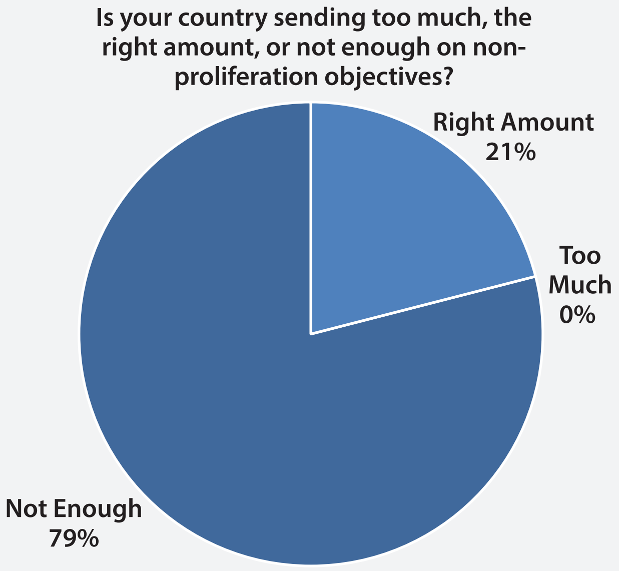
(15) In your opinion, have international non-proliferation efforts improved, stayed about the same, or regressed during the last year (2004)?

Responses: 84



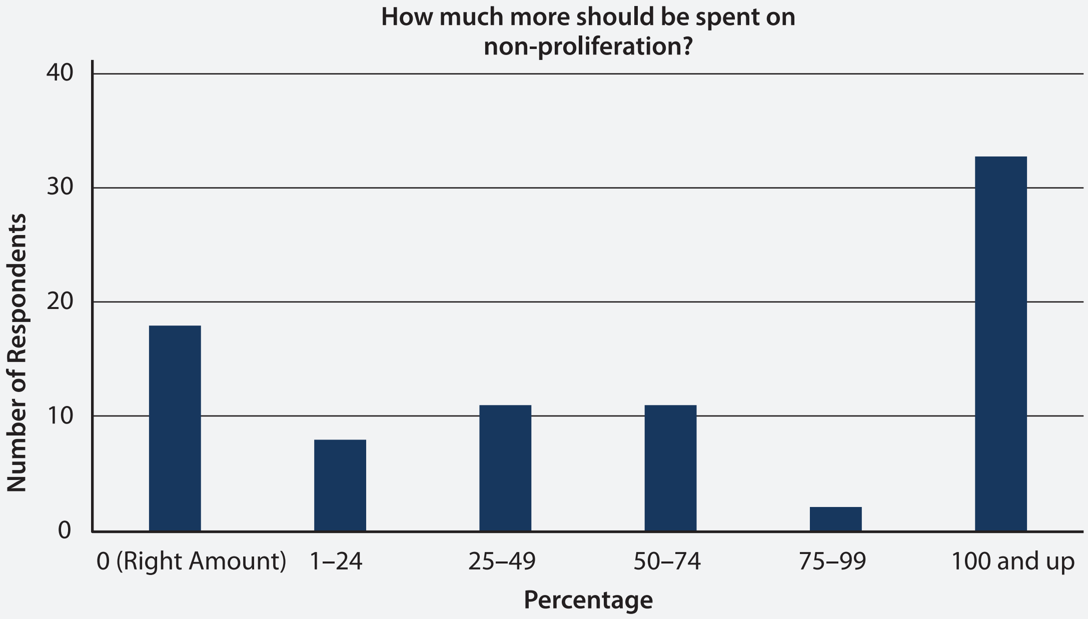
(16) Do you think your own country is spending too much, about the right amount, or not enough on non-proliferation objectives?

Responses: 84



(17) If you answered too much or not enough spending by your government, what percentage decrease or increase would you recommend?

Responses: 83 Median: 50%



(18) During the past year, a number of important steps were taken to enhance international non-proliferation cooperation. What do you regard as the most encouraging development that enhances global non-proliferation capabilities?

- Passage of UN Security Council Resolution 1540 on WMD proliferation?

  23 responses 15 exclusive responses

- Reaffirmation of the G-8 Global Partnership at Sea Island?

  12 responses 6 exclusive responses

- Expansion of the Proliferation Security Initiative?

  27 responses 20 exclusive responses

- Authorization of the first use of the Cooperative Threat Reduction program outside the former Soviet Union (to address chemical weapons in Albania)?

  20 responses 10 exclusive responses

- Formation of the Global Threat Reduction Initiative at the U.S. Department of Energy?

  14 responses 6 exclusive responses

- Other?

  Two responses cited the disruption of the A.Q. Khan network

  Respondents: 83

(19) In your opinion, what non-proliferation goal should receive the highest priority of the United States and the international community?

- Twenty-seven respondents cited the Nunn-Lugar Cooperative Threat Reduction.
- Fourteen respondents cited ending the nuclear programs of North Korea and Iran.
- Nine respondents cited worldwide control of fissile material. Three others cited controlling the nuclear fuel cycle.
- Eight respondents cited maintaining and strengthening the Nuclear Non-Proliferation Treaty.
- Four respondents cited Supporting and strengthening the administration's Proliferation Security Initiative to interdict illegal shipments of weapons and materials of mass destruction.
- Four respondents cited focusing on the proliferation threat from chemical and biological weapons.
- Four respondents cited rooting out the black market networks.
- Other suggestions for the top priority included implementing the Comprehensive Test Ban Treaty; implementing United Nations Security Council Resolution 1540 on WMD proliferation; developing sensors to detect smuggled nuclear material; developing better human intelligence on militant Islamic groups, and doing more to understand and counter the mindset of militant Islam; strengthening the International Atomic Energy Agency, the UN's nuclear watchdog; acknowledging Israel's possession of nuclear weapons; concentrating on the links between organized crime and proliferation; and protecting chemical plants near populated areas from terrorist attack.

(20) In your opinion, what proliferation risk or risks are most underrated or in the greatest need of additional resources or attention?

- Nine respondents cited the need for more effort to keep biological and chemical weapons out of terrorists' hands.
- Various respondents said that effort and funds should be devoted to Nunn-Lugar activities in the former Soviet Union, to Iran and North Korea, to nuclear disarmament, to controlling fissile material, and to controlling the fuel cycle.[8]
- One or more respondents cited WMD terror attacks not linked to Al Qaeda or militant Islam; threats to the food supply; nuclear material in Kazakhstan; a nondestructive but highly disruptive chemical or biological attack; the poor data available on the WMD technology base around and on WMD lethality; the security of nuclear weapons and materials in Pakistan; Russian tactical nuclear weapons; the preparedness of medical responders for a WMD attack on a city; the motivations for countries to seek nuclear weapons in the first place; weaknesses in the Proliferation Security Initiative and the export control regime; the intersections of criminal activity and terrorism; and the failure to match rhetoric with action.

(21) What studies or commentaries on non-proliferation issues that have appeared during the last year would you recommend?

- See the Lugar survey report for an alphabetized list of 29 recommendations.

[8] The number of respondents was not given.

Analysis of the Lugar Survey Results

The exploration herein primarily utilizes the data and information contained within the Lugar survey report itself. Additional information and understanding were provided by email with and a video interview with Dan Diller. Tracking and analyzing the citations and usage of the Lugar survey report also yielded valuable conclusions discussed in the section on uses and abuses of the Lugar survey. Finally, nearly twenty years have passed without occurrence of CBRN events—which provides additional data available for analysis.

When choosing participants for the Lugar survey, Lugar and his staff made effort to select experts across a broad spectrum of viewpoints and political leanings. That contributes to the sample being representative of the population of those in these subject areas. Asking a large number of experts (i.e., one hundred thirty-two) constitutes a sizable percentage of the entire population of such people. Having eighty-five out of one hundred thirty-two respond, with each answering a majority of the questions, is a reasonable response rate. However, focusing on "elites" rather than the larger pool of other nonproliferation experts limits the applicability of the results to only that "elite" subpopulation.

While the manner in which experts were chosen precludes statistically projecting the survey's results to the entire CBRN expert population, the knowledge and information acquired from the Lugar survey can be analyzed. Some interesting conclusions are provided herein, with the caveat that the analysis results that follow represent the subpopulation of "elites" in the time frame of 2005.

There were three major response types (i.e., response modes) to the Lugar survey questions:

(1) quantitative responses (questions 1–14 and 17);

(2) multiple-choice responses (questions 15, 16, and 18); and

(3) written responses (questions 19–21).

In terms of data analysis herein, most attention is paid to the first group. Limited categorical analysis can be done for the second group. Analysis is severely restricted for the third group from the few experts' comments provided. Yet, in terms of knowledge gained, the ordering is reversed. As discussed in the section on formal elicitation of expert knowledge topics, such a result does not have to hold. Formal elicitation methods capture the experts' thinking as they provide their answers, whether quantitative (numerical) or qualitative (multiple-choice and essay) responses are provided.

Providing Question Results

The Lugar survey report presented the question results in the following formats (reproduced in the Lugar Survey Questions subsection of this paper):

- Binned charts (resembling histograms)[9] were used to display quantitative responses from questions 1–5, 7, 9–14, and 17.
- Pie charts of percentages were used for questions 6, 8, 15, and 16.
- The tallies for the choices in question 18 were given but not displayed in a chart.

Likewise, response descriptions for some of the more frequently cited responses were documented for questions 19–21.

[9] Statistician Karl Pearson invented the graphic display of data called a histogram, which groups a set of data into bins or ranges of values plotted on the horizontal axis with the frequency or count of the number of data points in each group plotted on the vertical axis. Histograms function as a representation of the data's distribution and are particularly useful for discrete data such as the number of nations in questions 1–3 of the Lugar survey. Continuous data are discretized for a histogram representing the data's distribution; however, there should not be any gaps between bins to prevent loss of detail in the discretization. Common practice, especially in graphical software, employs equal-sized bins for continuous data; however, this may not always be the best choice for understanding the data's distribution. For example, the >0–9% bin used in the Lugar survey report masks very low responses of high interest.

The binned charts list the number of responses on the vertical axis and display bins for 10%, intervals, with 0% and 100% binned separately. Because 0% and 100% bins are listed separately, their neighboring bins are listed as >0–9% and 90–99%, respectively. Average and median values accompany the charts. Apparently, these statistics were calculated from the original responses of the experts, prior to binning, because attempts to reproduce the averages using the binned charts did not match the provided averages. Without the original responses, verification of the calculations of averages and medians is not possible.

Examining Conclusions Drawn from Numerical Results

Once an author states a conclusion in a document, such as the Lugar survey report, it tends to take on the role of a definitive, authentic, credible, quotable conclusion when read by others. However, without statistical analyses of the conclusion, in this case the survey results, it is unknown whether such a conclusion is warranted from the gathered data or whether it deserves such an elevated status.

For example, a number of statements made in the Lugar survey report compare responses to various questions. These comparisons are simple numerical comparisons of averages, without a statistical analysis that considers the broad distributions of responses. There is a difference between naively comparing average results and conducting a rigorous statistical/mathematical comparison. Lugar survey report conclusions about which answers were higher or lower than others cannot be justified without a statistical analysis to determine the significance of differences. The results of such analysis are given for the stated Lugar conclusions.

For those familiar with statistical analysis, each bullet below contains a footnote describing the test used and its level of significance, which is usually the choice of practice, 0.05 or 5%. For those who are unfamiliar, a brief explanation of the importance of determining significant difference is presented using the averages in the first bullet as an example.

Erroneous conclusions discussed in the results section of the Lugar survey report include the following.

- In discussing question 8, the Lugar survey report correctly observes that “a 55% majority of those responding (45 of 82) saw terrorist manufacture of a nuclear weapon after obtaining material as more likely, while 45% (37 of 82) believed that terrorist acquisition of a working nuclear weapon was the more probable scenario.”[10] While 55% is numerically larger than 45%, the statistical question becomes this: Is 55% significantly larger than 45%? For the 10% difference with eighty-two responses and a 5% significance level, there is no significant difference between 55% and 45%.[11] This may sound incredulous, so consider the following: if only nine people review a product, with five people liking it and four people not liking it, this 55% to 45% split is no different from half and half.

  Whether 55% is significantly different from 45% depends on three quantities: (1) the numerical difference of 10%; (2) the number of answers given to the question, which is eighty-two; and (3) the chance one is willing to accept that the stated conclusion is incorrect. This chance is chosen and is called the significance level, which is commonly 5% or 0.05. Changing any one of these quantities or a combination of the three can change a conclusion.

  (1) Suppose the numerical difference was 14% (57% versus 43%) with eighty-two respondents and a 5% significance level. For this increase in separation, the 57% is significantly larger than 43%.

[10] Lugar, *Lugar Survey*, 17.

[11] Binomial confidence intervals are used for the proportions to determine whether they contain the values of 0.45 and 0.55.

(2) Suppose ninety-two people had responded with the 55% to 45% split in answers. At the 5% level of significance, that 10% difference is now large enough to declare that 55% is significantly larger than 45%.

(3) Suppose the only change from the original is to increase the chance of being incorrect, choosing a 10% level of significance. Then 55% is significantly larger than 45%.

- On page 18, question 9, the probability of a biological attack in the next five years, with an average response of 19.7%, is stated to be "slightly more likely" than the probability of a nuclear attack, with an average response of 16.4%, for the same time period. One could argue the meaning of "slightly more," but the difference between these two averages is also statistically insignificant.[12] In other words, the difference between 19.7% and 16.4% is negligible for the responses to these questions.
- On page 21, question 12, the conclusion was that the average (30.5%) for a chemical attack was "lower than" the corresponding biological attack average (32.6%), both for the next ten years. Once again, while 30.5% is numerically less than 32.6%, the difference between these two averages is insignificant.[13] This is because the large variability in percentage estimates coming from a sample of approximately eighty experts is large enough to make 0.305 indistinguishable from 0.326.[14]

Of course, even naive numerical comparisons of averages and medians can be correct by chance. And, indeed, this is the case for the following conclusions:

- The following statement appears in the review of results in the Lugar survey report: "The median estimate of the probability of a radiological attack over ten years was twice as high as the estimate for a nuclear or biological attack during the same period."[15] This is numerically correct, and the median of 0.4 for the ten-year radiological attack is statistically larger than 0.2 for nuclear and biological ten-year attacks. In fact, that 0.4 is also significantly larger than medians of 0.1 (nuclear and biological five-year attacks) and 0.15 (chemical five-year and ten-year attacks).[16]
- In question 6, experts were asked whether they thought a nuclear attack in ten years was more likely to be perpetrated by a government or by terrorists. The pie chart shows that 21% of the eighty-five respondents chose a government and 79% chose terrorists. This split is statistically significant, making the terrorist scenario statistically more likely, as stated.[17]
- Similarly, for question 7, the conclusion that the black market is "the most likely" route to terrorist nuclear proliferation is statistically confirmed.[18]
- On page 11, question 2, "there was substantial agreement . . . on the number of new nuclear weapons states that would emerge." This

[12] The two-sided 95% binomial confidence intervals in Table 1 are used to determine whether the null hypothesis of equal percentages/proportions can be rejected.

[13] All statistical tests (conducted by Jane M. Booker) described herein used a 5% significance level, meaning there is a 5% chance that the stated conclusion is incorrect. Some analyses used JMP software from SAS Institute.

[14] The two-sided 95% binomial confidence intervals in Table 1 are used to determine whether the null hypothesis of equal percentages/proportions can be rejected.

[15] Lugar, *Lugar Survey*, 6.

[16] The 95% binomial confidence intervals are used for the medians (as proportions) to determine whether the null hypothesis of equal proportions can be rejected.

[17] The 95% binomial confidence intervals are used for the percentage as proportions to determine whether the null hypothesis of equal proportions can be rejected.

[18] Lugar, *Lugar Survey*, 16. The 95% binomial confidence intervals are used for the percentage as proportions to determine whether the null hypothesis of equal proportions can be rejected.

statement compares the number of nation states emerging in five years (question 1) with those in ten years (question 2). The question averages are 1.8 and 4.0, respectively, and the increase to 4.0 is significant over the 1.8 average despite the large variability of responses in question 2.[19]

Additional questionable conclusions include the following:

- On page 14, question 5, the conclusion is that "the respondents were much more pessimistic" on the risk of nuclear war within the next ten years than within five years. However, the average percentages between five and ten years (questions 4 and 5) are not statistically different; therefore, no increased pessimism exists. The large variations in the eighty-two responses to question 4 and the seventy-nine responses to question 5 are masked by only comparing the question averages (shown in Table 1). As discussed in the statistical analysis section below, pessimism can be concluded when taking the variability into account.[20]

There is at least one conclusion that cannot be tested because its meaning is uncertain. This statement is cited in an article by Carl Bialik,[21] listed as a citation example in the section in this paper on uses and abuses of the Lugar survey.

- The following statement appears on page 6 of the Lugar survey: "But the survey responses suggest that the estimated combined risk of a WMD attack over five years is as high as 50%. Over ten years this risk expands to as much as 70%." Yet, no combination of the averages or medians of the five- and ten-year combined WMD attacks produces values of 50% and 70%, respectively. If the binned charts for the four WMD five-year attacks are added together, there is a noticeable decrease in responses after the 50–59% bin and a similar decrease after the 70–79% bin for the combined ten-year attacks—a possible explanation. Such a prominently placed statement should include clarification on how the 50% and 70% values are determined from the responses.

As demonstrated in these examples, a common misconception in analyzing results from questionnaires is to state that the most popular answer, the one receiving the most responses, is the top result. The most frequent answer, however, is often not statistically different from others, and therefore, it is not the undisputed winner. Only results from statistical tests can determine when quantities are significantly different from one another. Conventional statistical tests for determining significant differences in survey questions include t-tests, tests for proportions, tests for averages, analysis of variance, correlations, and goodness-of-fit tests. We used many of these tests, as described in the section on statistical analyses of results, to determine which responses differ from others.

Displaying Numerical and Qualitative Results

For binned charts, such as those for questions 1–5, 7, 9–14, and 17, it is good practice to prominently list the average percentage and the median percentage along with the number of responses. These two central tendency statistics (average and median), along with the charts, demonstrate the asymmetric distributions of the responses.

Visual inspection of these binned charts indicates a wide variability in responses, extending across the entire range in many cases. To accompany that result, the standard deviation of responses would

[19] Tests of two means using a 0.05 level of significance indicate that the question 1 average is significantly smaller than the question 2 average.

[20] The two-sided 95% binomial confidence intervals in Table 1 are used to determine whether the null hypothesis of equal percentages/proportions can be rejected. It cannot.

[21] Bialik, "Pondering the Chances."

give a numerical value for comparison with what the eye determines.

For the pie charts for questions 6, 8, 15, and 16, it is likewise good practice to list the percentages in the pie slices and the corresponding number of responses.

For questions 18 and 19, the frequencies of the qualitative choices were provided and discussed.

Because of the open-ended nature of questions 20 and 21 (which was a good choice of response mode), statistical analysis is precluded unless it is possible to categorize the many and varied responses. For these questions, the valuable knowledge captured in these responses should have been preserved in its original form and made publicly available. Without the complete original written responses, it was not clear what, if anything, was learned from them. Did the experts say anything that was surprising or new? Did they say anything that was inconsistent with or contradictory to their other responses? Did they say anything that would indicate they did not understand the question or that would cause one to question their expertise?[22] Were strong biases evident?

Anticipating questions such as these, and subsequently addressing them, is part of the elicitation design. Probing questions can then be asked initially rather than as follow-up questions. Answers to these probing questions not only provide additional knowledge but can also provide the means for implementing analysis and uncovering insights.

Statistical Analyses of Results

Several statistical methods and their results are provided below for those interested in the details. For those who are not interested in statistical analyses, a summary of results is provided in the Conclusions from Our Analyses subsection in the section on analysis of the Lugar survey results.

As previously noted, the Lugar survey report displayed binned responses to many of the questions, accompanied by the average and the median. Those two quantities are statistics that characterize the central tendency of a distribution, like the binned charts. The other important characteristic of a distribution is the measure of dispersion. Examples of these include the standard deviation, the range, and specified quantiles (e.g., the twenty-fifth and seventy-fifth quantiles). None of those can be accurately determined from the binned charts because the bins contain a range of responses within them, except for the 0% and 100% bins. Because of this omission, statistical analyses methods are limited, and most of the results are based on the Lugar question averages, which by themselves do not represent the large variation of responses seen in the binned charts.

Even though the vast majority of experts answered all the questions, it is interesting to note whether any questions were avoided by many experts. Questions 6 and 19 were answered by all eighty-five. Question 3 stands out as having a statistically significant[23] larger number of nonresponses, with over a quarter of the experts not responding. This could be because respondents had more difficulty projecting twenty years into the future, and question 3 is the sole question asking for a twenty-year prediction.

The Lugar survey report provides a useful central tendency statistic, the average, for analysis of the quantitative questions. What, if any, differences exist among the question averages is the initial statistical question to be addressed. The binomial distribution and statistical inference (hypothesis)

[22] For example, one might question whether the expert who made the third comment on p. 34 knew about the nuclear fuel cycle.

[23] The phrase *statistically significant* or the word *significant* is used to indicate that a statistical test was completed and the result of that test demonstrates a difference in the quantities being tested, usually the average. The probability of the test producing an incorrect result is determined by the chosen significance level, usually 0.05 or 5%.

Table 1. Binomial Confidence Intervals for Average Percentages

Attack Type	Question Number	Prediction Interval (Years)	Number of Responses	Lugar Proportion Average	Lower Limit on 95% Confidence Interval for p	Upper Limit on 95% Confidence Interval for p
Nuclear	4	5	82	0.164	0.100	0.247
Biological	9	5	83	0.197	0.128	0.283
Chemical	11	5	83	0.201	0.131	0.280
Radiation	13	5	83	0.271	0.191	0.363
Nuclear	5	10	79	0.292	0.208	0.388
Biological	10	10	79	0.326	0.239	0.423
Chemical	12	10	80	0.305	0.220	0.401
Radiation	14	10	82	0.398	0.307	0.495

testing are useful in determining significant differences in the average percentages and probabilities provided. The binomial is also suitable for proportion testing for pie-chart questions.

Table 1 shows the averages for the questions covering the likelihood of different attacks in the next five and ten years in the Lugar survey. Questions 11, 12, 13, and 14 asked for probabilities, and questions 4, 5, 9, and 10 asked experts for probabilities expressed as percentages. Because of the mixing of definitions, percentages are divided by 100 in the fifth column of Table 1 and labeled as proportions.

The last two columns provide the upper and lower limits for the 95% confidence intervals from the binomial distribution for the averages in the fifth column. Calculation formulas for these confidence intervals are easily inserted into a spreadsheet that has the function for the inverse F distribution.[24]

The limits from the binomial confidence intervals represent a (95%) uncertainty bound for the true, but unknown, average proportion, p, for the population.[25] These limits form bounds on the averages (in column 5). If the interval from the upper and lower values (columns 6 and 7) in one row does not overlap the interval from another row, the averages for those two rows are significantly different (at the 5% level). If the interval from one row overlaps with the interval from another row, the averages of those rows cannot be considered significantly different.

The first result to notice in Table 1 is that all of the five-year intervals (the first four rows) overlap; therefore, they are statistically the same. The same is true of for the four rows showing the ten-year intervals. The nuclear, biological, and chemical five-year intervals are smaller than (do not overlap at all with) the radiation ten-year interval, making those five-year averages significantly smaller than the radiation ten-year average. Because the upper limit of the nuclear five-year average is so very close to the lower limit of the biological ten-year average, one could also conclude that these two averages significantly differ. Even though the Lugar five-year averages appear smaller than their ten-year counterparts, they are not significantly larger according to the row-by-row confidence interval comparisons in Table 1.

[24] Johnson and Litaker, "SAS® Program."

[25] A proper interpretation of a confidence interval is difficult to describe; however, it can be explained as follows: Let's say a 95% confidence for a probability is calculated for a sample of eighty. If one takes ninety-nine more such samples of eighty and calculates ninety-nine more such confidence intervals, then ninety-five of those intervals will contain the actual probability.

However, these results are based solely on the provided averages in the Lugar survey report, and they do not account for the visibly large variation (the spread of answers across the 0–100% scale) exhibited in all the quantitative binned charts. Accounting for that variation can change the results. Had the Lugar survey report provided all the individual numerical responses for each question, other statistical tests could determine whether the differences between the questions shown in Table 1 hold.

Another analysis (called analysis of variance) simultaneously examines the overall effects of four WMD categories with the effects of two time spans, five and ten years for the complete set of experts' responses. To demonstrate this, we estimated the individual responses from the chart bins. For this exercise, the bin midpoints and lower-end values are repeated according to the bin counts. Over the four WMD types and two time intervals, there are 651 estimated expert responses for the analysis.

The conclusions of this analysis of variance are as follows:

(1) Time intervals and WMD types are significant factors contributing to the variation in responses, and they provide some predictive capability for the responses.

(2) However, WMD type and time intervals only accounted for 8% of the total variability in the estimated individual responses. The remaining 92% variation in responses (as seen in the charts) is due to effects other than WMD type and time interval. It is not known what those factors are; however, if more information were available about the experts and their thinking and reasoning about their responses, additional explanations for the large variation could be possible.

(3) Not unexpectedly, the five-year average responses are significantly lower than the ten-year ones. The analysis of variance on the estimated 651 individual responses accounts for the large variability visually shown in the Lugar survey binned charts; therefore, this result is based on more information and is therefore more rigorous than results in Table 1, which compare question averages based on the binomial distribution.

(4) Analysis of variance results include comparisons among WMD types across time intervals. The comparisons in Table 1 are made without accounting for the large variation and compare WMD types given either the five-year or ten-year intervals. Despite this difference, the radiation estimates over both time intervals are significantly larger than those for nuclear, a consistent result from Table 1.

In the Lugar survey report, question 12 responses are described as "one of the most evenly dispersed set of responses."[26] This uniformity can be statistically tested by a chi-square (χ^2) goodness-of-fit test, which compares the distribution of the responses to that of a uniform distribution (where all responses occur with the same frequency). The χ^2 test indicates that the question 12 responses are *not* distributed uniformly. While none of the question response binned charts are uniform, question 14 responses are the closest to uniform.

A good way to demonstrate the dispersion (uncertainty) in the answers given by the experts for the numeric-response questions is by focusing on the tails (the responses with the largest and smallest values). For example, in question 1, the Lugar survey report specifically points out that four respondents believed five or more nations (the upper tail) would join, and five respondents selected the lower tail of zero. The experts' wide range of responses for question 1 reflects their diverse views. Some might conclude that this large range result (high uncertainty) is untenable; however, a large uncertainty result demands increased investigation to understand why experts had such different views.

[26] Lugar, *Lugar Survey*, 21.

Table 2. Test Results Indicating Large Counts in the 50–59 Bin

Attack Type	Prediction Interval (Years)	Is the 40–49% Bin Significantly Smaller than the 50–59% Bin?		Is the 60–69% Bin Significantly Smaller than the 50–59% Bin?	
Nuclear	5	Yes	1.22% < 6.10%	Yes	1.22% < 6.10%
Biological	5	Yes	3.61% < 8.43%	Yes	1.20% < 8.43%
Chemical	5	No	4.82% = 4.82%	Yes	1.20% < 4.82%
Radiation	5	Yes	4.82% < 14.46%	Yes	1.20% < 14.46%
Nuclear	10	Yes	1.27% < 16.46%	Yes	1.27% < 16.46%
Biological	10	No	8.86% ≮ 13.92%	Yes	3.80% < 13.92%
Chemical	10	Yes	5.00% < 13.75%	Yes	7.50% < 13.75%
Radiation	10	Yes	6.10% < 20.73%	Yes	1.22% < 20.73%

In the Lugar survey report, 47% of the experts selected the choice of “regressed during the last year.” This choice is statistically larger than the other two answers. However, it would have been interesting to analyze these responses by associating them with information about the experts, such as their country, areas or expertise or experience, and known viewpoints.

For the Lugar survey question 16, there is a slight discrepancy between the pie-chart percentage for “not enough” and that stated in the text. The chart shows 79%, and the text states “more than 78%;” the correct proportion is given as 78.6%, indicating a possible rounding difference. As expected, this “not enough” percentage is statistically larger than the other two choices.

The Lugar survey report binned charts for the responses to questions 4, 5, and 9–14 are collapsed into these bins: 0, >0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99, and 100. It is noticeable that the 50–59% bin has a large number of responses relative to its neighboring bins on either side. Could this be due to a large number of responses being 50%, which are absorbed into the 50–59% bins? It would have been informative to list the counts for the 50% response separately, as was done with the responses of 0% and 100%. Lacking this information, statistical tests can determine whether the counts in the 50–59% bins are larger than those in the 40–49% and 60–69% bins. Table 2 shows the counts for the 50–59% bins and their neighboring 40–49% and 60–69% bins. Columns 3 and 5 show whether the 50–59% bin has a statistically larger proportion than its neighbor for the five-year and ten-year questions, respectively. Combining all the rows gives 12.29% for the 50–59% bin, which is significantly larger than 4.45% in the 40–49% bin and the 3.69% in the 60–69% bin.

All except two neighboring bin comparisons show the unusually large number of responses in the 50–59% bins. This result suggests that many experts selected the 50% response. Had a separate bin for 50% responses been provided, its counts could be tested against its neighboring bins, as done in Table 2.

Insights and possible explanations for experts’ responses are not difficult to uncover with some additional information gathering and corresponding statistical analysis. For example, the names of the experts who responded are in the Lugar survey report appendix, and some information is known about them even without asking them for demographic or personal information. For example, it is known which ones are “scholars, policy makers, diplomats, and technicians.”[27] Comparison testing can be conducted to examine answers to questions according to which category each expert belongs

[27] Lugar, *Lugar Survey*, 4.

to. For example, do policymakers answer questions more pessimistically than scholars? Do experts from certain countries have similar views on their governments' spending (questions 16 and 17)?

The question-by-question results provided in the Lugar survey report group all the experts together for each question. However, by knowing which answers came from which expert (and this can be done by disguising the identification of experts, for example, as A, B, C, without citing their actual names), such data can permit analysis across questions for each expert. Results from such an analysis can answer questions like:

- Is each expert internally consistent?
- Does an expert tend to have a pessimistic or optimistic viewpoint?
- Is an expert not answering questions on a certain subject/topic?

While this type of exploratory data analysis usually requires a statistician, new conclusions about and understanding of the complete information content of the gathered knowledge is worth that effort.

It is not the goal of this paper to determine whether experts were correct or not in their estimates, given the lack of attacks five and ten years after 2005. However, it is interesting to note how the vast majority of Lugar survey experts did not provide answers at or near zero. There are statistical analyses for comparing experts' highly variable responses with the actual lack of attacks in the years since 2005.

For any application problem where the data are sparse and there are zero occurrences, the data can be enhanced by using prior existing knowledge, information, or relevant data and then combining it with the zeros. This is done using Bayes' theorem and is called Bayesian analysis.

Bayes' theorem involves three functions:

(1) The *prior* is a probability density function capturing the available information prior to obtaining data. A source like the responses from the Lugar survey can take the role of the prior information.

(2) The *likelihood* is a function that is formed from the data. In this case, the likelihood is the number of attacks (including zero) for the elapsed five-year and ten-year time periods in the questions.

(3) Mathematically combining the prior with the likelihood is accomplished using Bayes' theorem. That formula produces the combined result called the *posterior* distribution.

Caution is required when selecting the prior and the likelihood because these choices influence the posterior result. For example, a strong prior can overpower the data, which are usually sparse, and the prior and the likelihood may not overlap, which produces the posterior for values that are present neither in the prior nor the data. Although there are many other difficulties in doing Bayesian analysis, there are two major advantages: (1) when the data are sparse, prior information increases the information content when combined with the data, and (2) such a combination reduces variability. Because of the way the Lugar survey was constructed and executed, any prior formulated from it would be of questionable quality. For purposes of illustration only, using the Lugar responses as a prior with the elapsed time data for the likelihood only reduces[28] the average from 16.4% to 12.8% in question 4 and from 29.2% to 23.2% for question 5.

Besides Bayesian analysis, there are other ways to analyze the predictive capability of the Lugar survey experts by using the data of one ten-year time lapse and three five-year time intervals since 2005. For illustration purposes, Tables 3 and 4 provide examples showing how poor the predictive capability is using Lugar survey results. Had efforts been made to capture the experts' thinking and

[28] The entire Lugar survey binned chart was fit to a beta distribution and the zero attacks data followed a binomial distribution for this Bayesian analysis. This was done for questions 4 (five-year attack) and 5 (ten-year attack).

problem-solving when answering these questions, their poor performance may have been understood. Had ranges of responses been asked, their collective performance may have been better.

Because many of the results of the numerical-response questions were binned in the same manner, the chosen bins themselves can be analyzed rather than the collapsed data in them. Given that nearly all bins contained some experts' responses for the WMD questions, the question becomes: What are the probabilities of zero attacks in one, two, three, and four elapsed time intervals for each bin? In other words, if the experts were good predictors, what probabilities should they have estimated to correspond to zero attacks in five years up to twenty years?

Table 3 shows the calculated probabilities, using the binomial distribution, for the time-elapsed data of zero attacks in the rightmost four columns. The middle of each bin was rounded to the nearest 5% integer value in each Lugar bin, as shown in the second column. Those percentages were converted to probabilities, p, in the third column and used in the binomial distribution to calculate the probability values in the rightmost four columns.

This analysis does not involve any expert responses—the bins' middle integer values represent those experts who provided responses in that bin for any question. The probabilities of zero attacks in the rightmost four columns represent values near to what the experts should have provided if they had good predictive capability. For example, the probability of zero attacks in one elapsed time interval (column 4) is 0.95 for experts answering >0–9% for questions 4, 9, 11, and 13. The probability of zero attacks in two elapsed time intervals (column 5) is 0.90 for experts answering >0–9% for questions 5, 10, 12, and 14.

The elapsed time intervals in the last four columns require explanation because the definition of the time interval is either five years or ten years. The probabilities given in column 4, for one time interval, apply to one five-year elapsed time interval and also for one ten-year elapsed time interval. Likewise, the probabilities in column 5 apply to two five-year intervals and two ten-year intervals. By 2025, all these elapsed intervals will have occurred since 2005. However, the last two columns have also elapsed for three and four time intervals only for responses to the five-year questions. According

Table 3. Binomial Probability of Zero Attacks Using Bin Middle Integers

Bin	Bin Middle Integer	Bin Middle p	Probability			
			Zero Attacks in One Time Interval	Zero Attacks in Two Time Intervals	Zero Attacks in Three Time Intervals	Zero Attacks in Four Time Intervals
0	0%	0	1	1	1	1
>0–9	5%	0.05	0.95	0.90	0.86	0.81
10–19	15%	0.15	0.85	0.72	0.62	0.52
20–29	25%	0.25	0.75	0.56	0.43	0.32
30–39	35%	0.35	0.65	0.42	0.28	0.18
40–49	45%	0.45	0.55	0.30	0.17	0.09
50–59	55%	0.55	0.45	0.20	0.09	0.04
60–69	65%	0.65	0.35	0.12	0.04	0.02
70–79	75%	0.75	0.25	0.06	0.016	0.004
80–89	85%	0.85	0.15	0.02	0.003	0.0005
90–99	95%	0.95	0.05	0.003	0.0001	0.000006
100	100%	1	0	0	0	0

Table 4. Prediction Capability Using Poisson Distribution

Attack Type	Lugar Average	Probability of Zero Attacks	0% Bin Correct	Bins 0–20% Counts	Bins 0–20% Correct	Zero Attacks in Ten Years
Nuclear five-year	0.164	0.61	0.05	56	0.68	0.72
Biological five-year	0.197	0.53	0.04	47	0.44	0.67
Chemical five-year	0.201	0.55	0.05	46	0.57	0.67
Radiation five-year	0.271	0.44	0.02	33	0.38	0.58
Five-year total	0.208	0.54	0.04	182	0.55	0.66
Nuclear ten-year	0.292	0.75	0.01	35	0.55	
Biological ten-year	0.326	0.72	0.01	30	0.44	
Chemical ten-year	0.305	0.82	0.04	35	0.40	
Radiation ten-year	0.398	0.82	0.02	23	0.28	
Ten-year total	0.331	0.76	0.02	123	0.38	

to the last column, for good predictability, the vast majority of experts should not have provided answers in any bin beyond 10%. Looking at the averages in Table 1, their predictability is poor given the data of zero attacks for these WMD types.

Table 4 shows a direct comparison of the averages for the Lugar survey WMD responses to probabilities of zero attacks that occurred in the five-year and ten-year elapsed time intervals. While using the question averages ignores the high variability evident in the binned charts, the average is a measure of the central tendency of the collection of experts responding.

For this analysis, the Poisson distribution is used to calculate probabilities of attacks using a specified failure rate applied to the number of elapsed intervals (e.g., three five-year intervals and one ten-year interval). The Lugar survey averages serve as the attack rate Poisson parameter for each question. Column 3 is the probability of zero attacks from the Poisson distribution, which are all about 50% based on experts' five-year estimates. This means that using the experts' averages to define the attack rate is no better than flipping a coin. The Poisson probabilities for zero attacks in the single elapsed ten-year time interval are markedly higher based on the experts' higher estimates. The rows labeled total contain the summations over the four WMD questions. Notice there is little variation among these four WMD types in both the five-year and ten-year questions.

The fourth column of Table 4 lists the proportion of experts who estimated zero attacks. Comparing these to the Poisson probabilities (third column) shows how poor their average predictability is.

To illustrate the benefit of asking experts to provide a range of values corresponding to their uncertainty about a single-valued response, assume experts provided ranges that spanned from 0% to 19%. The counts and proportions from collapsing those three Lugar chart bins are in the fifth and sixth columns of Table 4. While the fourth column, the 0% bin only, indicates poor expert predictive ability, the sixth column values correspond better to the Poisson probabilities for zero attacks given in column 3, indicating better predictive capability than the single estimate of zero. This example illustrates the importance of asking experts to provide their uncertainties, which can be easily done by asking them to specify a range of values around their single-valued response. The purpose is not necessarily for them to hedge their bets, but to allow them to better express their expertise, knowledge, and uncertainty.

Finally, Table 4 addresses the question of what the ten-year prediction would be using the experts' average five-year failure rate (second column) for the Poisson. The last column of Table 4 lists the Poisson probabilities for two elapsed five-year time intervals. One would expect that these values would align with the experts' ten-year attack estimates, if the experts' provided realistic increased ten-year estimates. However, the experts underestimated their ten-year values given their five-year estimates.

Conclusions from Our Analyses

The Lugar survey report included several conclusions drawn from visual inspection of the responses to the survey questions. These statements are statistically tested. Many are not valid, as discussed in the section of this paper on examining conclusions drawn from numerical results.

Our analyses reveal (refer to Table 1) that the nuclear, biological, and chemical five-year averages are significantly smaller than the radiation ten-year averages, and the nuclear five-year is smaller than the biological ten-year average. Even though the Lugar five-year averages are numerically smaller than their ten-year counterparts, those differences are not statistically significant. This result demonstrates the need for statistical analysis instead of visual interpretation.

In addition, our analyses of variance on the complete set of 651 estimated responses concludes that the five-year responses are significantly smaller than the ten-year ones. In addition, the WMD types and time intervals are both significant contributors predicting responses; however, these account for only a small fraction of the total variability in all responses. More information is needed to explain the responses. Additional analyses and conclusions would have been possible had the Lugar survey report listed all individual responses for each question, without identifying experts by name.

No reason was offered for why more than a quarter of the experts did not answer question 3.

The large variation in the binned charts indicates large uncertainty in the experts' cognition and problem-solving. That uncertainty is also indicated by the significantly large counts in the 50–59% bins in Table 2. Wide variability demands further investigation to understand its source(s), and that requires additional information from the experts as to their thinking and problem-solving when answering questions.

Contrary to the conclusions stated in the Lugar survey report, the distributions of responses are not uniform for any of the questions—a result that is not unexpected.

Our analyses of experts' predictive capability illustrate how the Lugar survey responses compare with the elapsed time intervals since 2005 with zero WMD attacks occurring. Analyses demonstrate the experts' poor predictability. However, had they been given the opportunity to provide information on their thinking and problem-solving while answering the questions, their poor predictability might at least be understood. For example, the simple addition of asking experts for ranges around their answer would have given better insight into why they appeared to be poor predictors. Our examination of experts' predictive capability is only meant to illustrate the importance of designing a survey or elicitation to capture experts' problem-solving, thinking, knowledge, and uncertainties.

A final caution is that the above analyses apply solely to the Lugar survey responses and are not applicable to the entire population of the nonproliferation knowledge in 2005 (or any other year). Inferring anything about the whole population of nonproliferation experts from the Lugar survey respondents is not justifiable given the way the experts were chosen, how the survey instrument was designed and executed, and the purpose of the survey.

Uses and Abuses of the Lugar Survey

In introductory remarks to the Lugar survey report, Senator Lugar acknowledges a critical limitation of the survey's methodology, defines his intent, and describes the anticipated applicability of the report's findings:

> I would underscore that this study is not meant to be a scientific poll of the entire national security community. Rather, my intent was to discover consistencies and divergences in attitudes about non-proliferation among a large and diverse group of well-informed experts. . . . I believe that the results of this survey will be useful in helping to define the parameters of the risks that we face, assessing the current state of non-proliferation and counter-proliferation efforts, and identifying issues of concern that require more attention. I am hopeful that it will provide a point of reference for scholars and practitioners, as well as those who do not follow proliferation issues on a daily basis.[29]

These sentiments were more recently echoed by Dan Diller,[30] former legislative director for Senator Lugar and an organizer of the Lugar survey, who affirmed that it was not intended to be a scientific evaluation of the risks of WMD use, but instead a means to prompt discussion on the importance of nonproliferation. He explained that the Lugar survey sampled the opinions of a select group of experts to better understand the spectrum of WMD threats, raise awareness among the policy community and the public, and prompt further research and discussion through documents published by news sources, policymakers, and academics.

The Lugar survey was certainly successful in its attempt to motivate further discussion. We conducted a literature search of documents referring to the Lugar survey from 2005 to mid-2023, the time of this writing. Based on this literature search, we compiled 119 documents that have referenced the survey over the last almost two decades. We found these documents through open-source searches on databases including Google Scholar, ProQuest, and JSTOR. We also found documents through general searches on Google, Bing, and YouTube. While we do not believe this search uncovered every document that cited the Lugar survey, we are confident that our search found the large majority of available open-source documents. Such documents include academic articles,[31] research papers,[32] newspaper pieces,[33] blog posts,[34] and congressional hearing materials.[35] A database summarizing these documents and highlighting unedited references to the Lugar survey is available in the online appendix to this paper.

Figure 1 provides a histogram of documents citing the Lugar survey over time. Year 2023 has only a partial count as it does not include documents published after mid-2023. Four documents do not record the publication date, so they are not included in this chart. Documents are categorized according to their overall perspective on the Lugar survey as follows:

- **Negative** comprises documents that explicitly criticize the methodology or results of the Lugar survey.
- **Questioning** is assigned to documents that express some concern(s) regarding the methodology or the respondents' potential biases, but use Lugar survey values in their analysis nonetheless.

[29] Lugar, *Lugar Survey*, 4.

[30] Interview with Dan Diller, director of policy at the Lugar Center, on August 11, 2022.

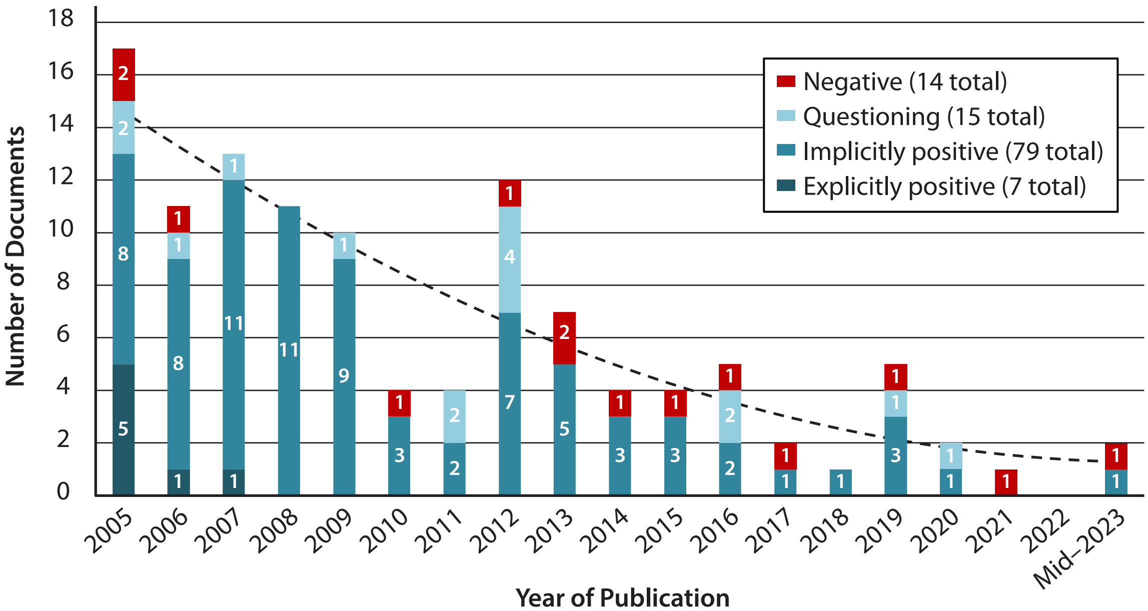
[31] For example, Bunn, "Mathematical Model."

[32] For example, Kunreuther and Michel-Kerjan, "Terrorism Risk Financing Solutions."

[33] For example, Bialik, "Pondering the Chances."

[34] For example, Gongol, "Structural Failures."

[35] For example, *Trends in Illicit Movement of Nuclear Materials*.



We stopped searching for documents at the end of June 2023; thus, all documents found are from 2005 through mid-2023. Four undated documents are not included in this chart.

Figure 1. Documents Citing the Lugar Survey from 2005 through Mid-2023

- **Implicitly positive** refers to documents that reference the Lugar survey's results with no discussion or criticism of its methodology. Documents in this category include those that reference the Lugar survey only in a footnote or in sections on further reading or sources consulted.
- **Explicitly positive** consists of documents that praise the Lugar survey, including blog posts and news reports dedicated solely to discussion of the Lugar survey's findings.

We define the first category as "critical" and the last three categories as "uncritical" hereinafter.

Figure 1 displays the counts of documents in each of these categories over time. The legend in Figure 1 notes the totals in each of the four categories. We have also drawn a quadratic fit[36] to the annual totals to guide the eye in assessing the trend in numbers of documents over time.

Our database of these documents includes title, author(s), publication year, publication and publisher, overall perspective on the Lugar survey using the categories defined above, and Lugar survey questions referenced in the document. Two examples of database entries appear on the following page. Details on all documents, in this format, can be found in the online appendix.

The most significant findings from our literature analysis are summarized below. These findings underscore the ease with which Lugar's nonscientific results have been used and abused by experts and nonexperts alike.

The majority of documents that reference the Lugar survey are uncritical of its methodology or results. Since its publication, the Lugar survey has been treated as a scientific study through the continual misinterpretation of the validity of its findings.

[36] Quadratic fit line: $y = 0.0394x^2 - 1.5273x + 16.199$

10	**Title:** Pondering the Chances of a Nuclear Attack	**Author(s):** Carl Bialik	**Year:** 2005
	Publication (Publisher): News article (*Wall Street Journal*)	**Perspective on the Lugar Survey:** Negative	**Survey Question(s) Referenced:** 5, 6, 14, 20
	Link: https://www.wsj.com/articles/SB112059629605777656		
	Quote: "But how do you predict the likelihood of an event that has never happened before? "The past is the baseline for predicting the future. In forecasting company revenue, economic indicators and hurricane counts, experts start with prior numbers and adjust them higher or lower to reflect expected future trends. When it comes to estimating the chance of a terrorist attack using biological or nuclear weapons, it's hard to go beyond an educated guess. "Two weeks ago, Sen. Richard Lugar (R., Ind.), chairman of the Senate Foreign Relations Committee, released the results of an ambitious survey of arms experts. The study was conducted in late 2004 and early 2005. On average, the 85 respondents predicted a 29.2% chance of a nuclear attack in the next decade, with 79% saying that such an attack was more likely to be carried out by terrorists than by a government. Sen. Lugar said in the report that 'the estimated combined risk of a WMD attack over five years is as high as 50%. Over 10 years this risk expands to as much as 70%.' . . . "Yet there are also drawbacks. As well-informed as arms experts are, and as well-intentioned, I'd argue they have a natural bias toward overstating risk — greater risk increases the value of their expertise, and, therefore, their prominence and even funding. Politicians who commission such predictions likely do so because they want to raise awareness, a goal best served by alarming results."		
18	**Title:** "Dirty Bomb" Attack: Assessing New York City's Level of Preparedness from a First Responder's Perspective	**Author(s):** John Sudnik	**Year:** 2006
	Publication (Publisher): Thesis (Naval Postgraduate School)	**Perspective on the Lugar Survey:** Explicitly positive	**Survey Question(s) Referenced:** 13, 14
	Link: https://apps.dtic.mil/sti/citations/ADA445265		
	Quote: "Perhaps the most compelling case made for the probability of an RDD attack is put forth in a 2005 survey conducted by U.S. Senator Richard G. Lugar, chairman of the Senate Foreign Relations Committee. The survey polled a group of leading national security experts on various WMD proliferation issues. In comparison to the threat of a chemical, biological, or nuclear attack on a major city, the survey group found . . . the risk of a radiological attack as significantly higher. The median and average estimates of risk were 25% and 27.1% respectively over the next five years. Over ten years, both the median and the average estimate of risk jumped to 40%. The median estimate of the probability of a radiological attack over ten years was twice as high as the estimate for a nuclear or biological attack during the same period."[37]		

The Lugar survey and its results have been referenced in an expansive variety of documents, including those authored by researchers, students, professors, news reporters, policymakers, and scientists. In contrast to the Lugar survey report itself, however, most of these documents do not acknowledge its limitations; rather, contrary to the clear caveat in the survey report, they use the results of the survey as scientific findings. Of the one hundred and nineteen documents we identified that cite the Lugar survey, one hundred and five (~88%) fall within the categories we characterize as uncritical. Of the one hundred and fifteen dated documents, one hundred and one are uncritical. Authors of these documents use Lugar survey values with little or no questioning of the survey's methodology or the representation of its results.

Within the category of "uncritical," fifteen documents are labeled as "questioning," because they point out the problem of biases of survey respondents or otherwise question the survey's methodology but use the survey's results nonetheless. For example, in his report *Safeguarding the Future: Cause Area Report*, John Halstead cautions that the Lugar survey and similar elicitation attempts are "likely subject to significant subject bias and selection effects, but at least suggest that the risk is non-negligible."[38] Niyazi Onur Bakir and Detlof

[37] The bulleted points in the Examining Conclusions Drawn from Numerical Results subsection of this paper contain the statistical evaluation of this and other Lugar survey report statements.

[38] Halstead, *Safeguarding the Future*, 37.

von Winterfeldt, in "Is Better Nuclear Detection Capability Justified?," use the Lugar survey results to explain the risk of radiological attack, noting that "while these numbers are probably too high due to common biases in probability estimation, they reflect concerns based on evidence."[39]

The majority of documents (eighty-three of one hundred and nineteen total documents) citing the Lugar survey reference its results without caveats and are thus categorized as "implicitly positive." As such, they neither applaud its elicitation practices or results nor discuss its flaws. This amounts to an implicit endorsement of the survey. For example, in their report, *Use of Nuclear and Radiological Weapons by Terrorists?*, Christoph Wirz and Emmanuel Egger use Lugar survey results to conclude that there are no significant obstacles for terrorist organizations to acquire WMD. They reference the responses to Lugar survey question 14, the probability of radiological attack in ten years, and claim that the estimated 40% median probability of attack is reason to increase disaster and readiness preparation for first responders and public educators.[40] As another example, Roland Schenkel, in his book chapter "Improving Verification: Trends and Perspectives for Research," cites the Lugar survey briefly, noting "a recent survey issued by Lugar about the possibilities of an attack based on nuclear, biological or chemical weapons shows that there is a real risk."[41] Although he references no specific questions, he uses this Lugar survey result to highlight the increasing threat of terrorist use of WMD.

Finally, some documents explicitly commend the methodology of the Lugar survey or use its results for parameters of their derivative analyses. Seven documents explicitly condone the Lugar survey and are thus labeled "explicitly positive." One such document, a thesis titled "'Dirty Bomb' Attack: Assessing New York City's Level of Preparedness from a First Responder's Perspective" by John Sudnik, calls the Lugar survey the "most compelling case made for the probability of an RDD [radiological dispersion device] attack."[42] Another document, *Securing the Bomb 2007* by Matthew Bunn, uses the Lugar survey as a prime example of eliciting "well-informed analysts" on the risk of nuclear terrorist attacks.[43] Four documents solely detail the findings of the Lugar survey and provide no caveats or questioning of Lugar's methodology.[44]

In stark contrast to these one hundred and five total uncritical documents, only fourteen documents explicitly discuss flaws in the methodology of the Lugar survey and are thus labeled "negative." Many of these documents criticize the results by pointing to respondents' natural biases toward overstating WMD threats. For example, Michael Huemer, in his book *The Problem of Political Authority*, cautions that, although these assessments may appear useful, they "should be taken with a grain of salt, as national security experts may have a bias toward overstating threats to national security. Those who are most predisposed toward concern about national security threats are most likely to become national security experts."[45] Carl Bialik also discusses this bias in his *Wall Street Journal* article "Pondering the Chances of a Nuclear Attack," claiming that WMD and national security experts "have a natural bias toward overstating risk—greater risk increases the value of their expertise, and, therefore, their prominence and even funding."[46]

[39] Onur Bakir and von Winterfeldt, "Better Nuclear Weapon Detection Capability," 1.

[40] Wirz and Egger, *Use of Nuclear and Radiological Weapons*, 508.

[41] Schenkel, "Improving Verification," 592.

[42] Sudnik, "'Dirty Bomb' Attack," 21.

[43] Bunn, *Securing the Bomb*, 42. Interestingly, Bunn's paper *Controlling Nuclear Warheads and Materials: A Report Card and Action Plan* was featured in the 21st question in the Lugar survey on recommended nonproliferation studies and commentaries.

[44] "Sen. Lugar Releases New Report"; Digges, "US Survey"; Chapman, "New Report Paints Grim Picture"; and "Experts Assess Likelihood."

[45] Huemer, *Problem of Political Authority*, 311.

[46] Bialik, "Pondering the Chances."

Other "negative" documents focus on inherent flaws in attempting to quantify highly uncertain and multifaceted WMD risks in the first place. In *Minimum Deterrence: U.S. Nuclear Weapons and the Priority Threat Facing the United States*, the National Institute for Public Policy aptly describes the limitations in such estimates: "the inherent problem with quantifying the probability of such complex human actions with this type of precision is that the knowledge required to make these claims credibly spans the areas of psychology, sociology, history, physics, chance, and unknown/unknowable factors that can affect the system under study."[47] In a congressional hearing on the illicit movement of nuclear materials, Raymond Juzaitis, then associate director at the Lawrence Livermore National Laboratory, makes a similar statement, arguing that "there are too many human factors involved" to make the decisive mathematical assessments found in the Lugar survey.[48] These two documents address the fundamental challenge with attempting to quantify the risk of a complex, uncertain, and unpredictable threat.

The wide range of responses to most questions is one of the Lugar survey's most interesting results; however, results from the Lugar survey other than median and average values are rarely cited.

The Lugar survey report includes a sizable amount of information on respondents' answers to likelihood questions, including the number of responses; a binned chart that visually depicts the mode and the range of responses, as well as the variability of responses; the calculated average and median responses; and descriptive text on the survey responses. The mode, average, and median are measures of central tendency of the responses. While interesting, they do not capture important information about the variability in responses, which for many questions is simply striking and the most significant observation. Since the average and median are the only presented calculated values, it is not surprising that documents that cite the survey focus on these values as the most important results. Lugar and his staff clearly emphasized these values, as they anticipated they would garner more press attention.[49]

The Lugar survey report represented the variability in participants' responses visually through bar charts, rather than calculating the variance (or standard deviation) in responses to each question. Thus, as there is no simple representation provided, it is not surprising that fewer citing documents discuss variability. Of the one hundred and nineteen documents evaluated, ninety-four specifically cite values from the Lugar survey. Of those ninety-four, thirty-seven cite the median or average, while only fourteen mention the variability of responses.

Quantitative estimations of nuclear risks are most cited in the literature, followed by radiological risks. Qualitative responses are seldom cited, contrary to the survey organizers' view that these would be the most informative.[50]

The Lugar survey is divided into two parts: (I) Assessing Proliferation Threats and (II) International Non-Proliferation Responses. Part I records and bins risk estimates into quantitative results, whereas the part II records and synthesizes written prose into qualitative results. Within part I, no question results were significantly different statistically, but some results were cited in documents more frequently than others.

Figure 2 shows the breakdown of the total number of documents in our literature database that cite

[47] National Institute for Public Policy, *Minimum Deterrence*, 11.

[48] *Trends in Illicit Movement of Nuclear Materials.*

[49] According to Dan Diller, the final write-up of the Lugar survey focused on the final risk percentages in the first portion because they knew that these would get the most news attention. The second section, where the survey focused on qualitative responses to questions on nonproliferation efforts, was designed to provide better understanding of international nonproliferation priorities for policymakers and academics.

[50] Conversation with Dan Diller, director of policy at the Lugar Center, on August 11, 2022.

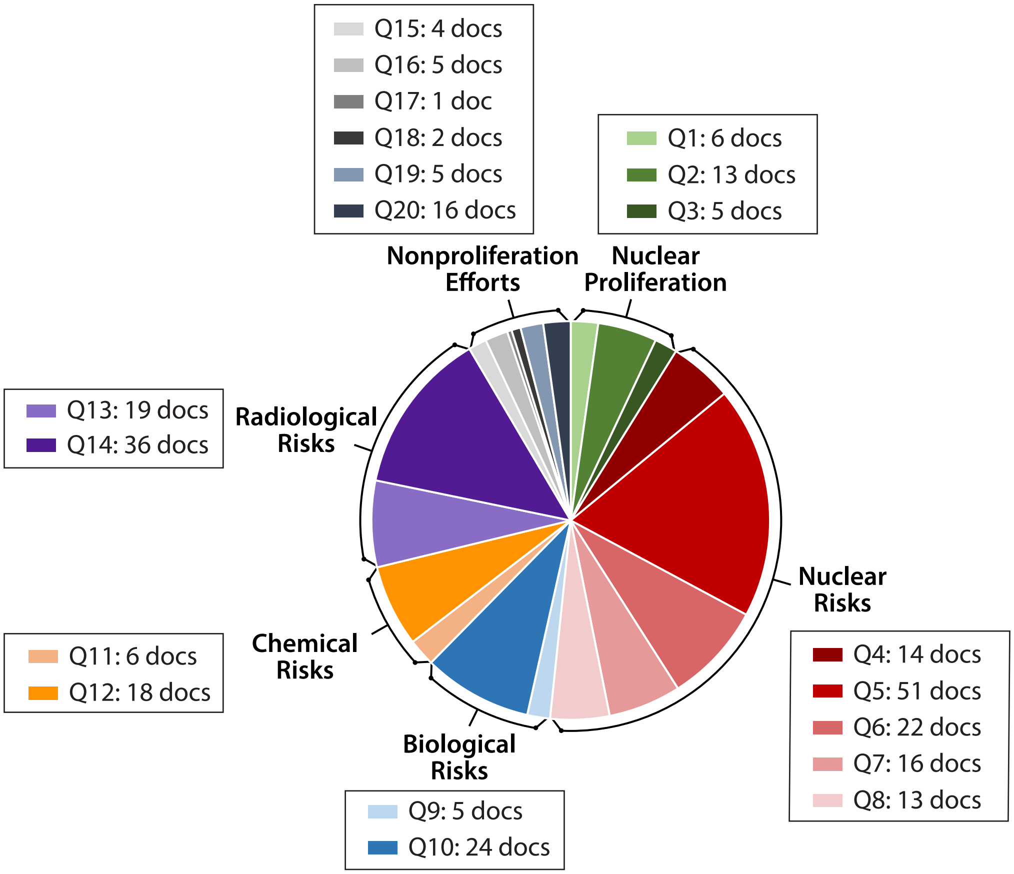


Figure 2. Number of Documents That Cite Each Question in the Lugar Survey

each question in the Lugar survey. Question citations were tallied for every document that mentioned them. Since many documents reference multiple questions, the total number of questions counted is greater than the total number of documents. The most referenced questions were on nuclear attack risk (questions 4–8), followed by radiological attack risk (questions 13 and 14) and biological attack risk (questions 9 and 10).

Questions on nonproliferation efforts (part II of the survey) received the least attention in the literature. Although 30% of the questions in the survey are from part II, only 9.5% of questions referred to in the literature are from this part of the report. Unlike part I, part II asked respondents for open-ended answers to broad questions, allowing for more detailed responses. The topics for questions in part II were "status of international non-proliferation efforts," "government spending on non-proliferation programs," "recommended spending increases," "encouraging developments in non-proliferation capabilities," "non-proliferation priorities," and "underrated non-proliferation risks." According to Dan Diller, responses to these questions provided a more accurate and detailed account of survey respondents' beliefs on nonproliferation.

Over one in four citing documents identified Lugar survey findings on the risk of radiological attack as the most insightful results from the survey.

A significant portion of documents (thirty-eight of one hundred and nineteen) specifically reference the Lugar survey findings on the likelihood of a radiological weapon attack. The median risk of radiological attack for the next five years was 25% and the average was 27.1%. Both the median and average were approximately 40% for the next ten years.[51]

Many of the documents citing these results focus their analyses and findings on the high threat of radiological attack. For example, the Defense Science Board, in its report *Reducing Vulnerabilities to Weapons of Mass Destruction*, characterizes the Lugar survey findings on the risk of radiological attack as "somewhat surprising," saying that "securing radiological material everywhere in the world should be a high priority if one agrees with these experts."[52] In "The Economics of Nuclear Energy Markets and the Future of International Security," Erwann Michel-Kerjan and Debra K. Decker use a potential radiological attack to demonstrate the risk of terrorist acquisition of nuclear weapons. They use the Lugar survey's findings to raise concerns, explaining that "fears surround the spread of nuclear energy and the possible diversion of nuclear materials from the fuel cycle process."[53]

Since Lugar survey respondents identified radiological attack as the most likely form of WMD attack in both five- and ten-year projections, the Lugar survey provides a convenient point of validation for these studies. We have not found any documents emphasizing a low risk of radiological attack that cite the Lugar survey as a counter.

We have uncovered little direct evidence that the Lugar survey has been instrumental in changing minds or in developing nonproliferation policy.

The Lugar survey's goal was to raise public awareness and promote public debate on nonproliferation issues. Despite this intention, the survey and its results have been rarely cited in federal policy documents or congressional hearings. In our open-source search, we uncovered no official legislative or executive branch policy documents that reference the Lugar survey or its results. Of all citing documents, only three are federal documents: two hearings before the Committee on Homeland Security[54] and a report from the Department of Homeland Security on the National Small Vessel Security Summit.[55] Additionally, one Congressional Research Service report prepared for members and committees of Congress mentions the Lugar survey and its results.[56]

In addition, three municipal- or state-level governments used the Lugar survey to promote disaster relief funding. The University at Buffalo published the presentation "A Preliminary Multihazard Risk Profile for New York State,"[57] and the California Department of Public Health published the presentation "Quantifying Unfortunate Events for Strategic Planning and Resource Allocation."[58] Both documents used the Lugar survey to introduce the risk of disaster and advocate for new disaster funding and policy. Additionally, one city in Alaska—Ketchikan—published a brief description of the survey with no apparent funding or policy motive.[59] We found no published responses to these presentations or reports.

[51] Table 1 shows the Lugar survey averages and statistically compares them. Even though the five-year radiological average appears to be the largest, it is not statistically larger than the other three averages. The same is true for the ten-year averages. Only the ten-year radiological average is larger than the five-year nuclear, biological, and chemical averages.

[52] Defense Science Board, *Reducing Vulnerabilities*, 12.

[53] Michel-Kerjan and Decker, "Economics of Nuclear Energy Markets," 28.

[54] *Trends in Illicit Movement of Nuclear Materials*; and *Review of U.S. International Efforts*.

[55] Brownstein et al., *Report of the DHS Small Vessel Security Institute*.

[56] In an interview, Dan Diller alluded to other congressional testimonies that may have referenced the Lugar survey, but he was unsure where to find them. They were not found in our open-source search.

[57] Allen et al., "Preliminary Multihazard Risk Profile."

[58] Anderson and Thomas, "Quantifying Unfortunate Events."

[59] "Experts Assess Likelihood."

Thus, although Lugar intended to use the survey to inform policy, the survey's results are rarely referenced to support nonproliferation legislation or federal discussion of nuclear issues.

Conclusions

In this subsection, we present our conclusions regarding uses and abuses of the Lugar survey. In the context of the survey's diminishing relevance over time, we focus on the responsibilities of the survey team, survey participants, and citing authors in avoiding the major abuses in future similar surveys.

Role over time. Not surprisingly, as evident in Figure 1, documents referencing the Lugar survey gradually diminished in number after the survey report was released in 2005. In the first decade after its publication, the report appears to have played a noteworthy role in media and expert portrayals of the probabilities of a WMD attack (especially a nuclear attack), terrorist acquisition of nuclear weapons, and nuclear proliferation. At the same time, we have uncovered little tangible evidence that it changed any minds; rather, its use seems more consistent with reinforcing experts' and policymakers' existing perceptions. In any event, the extent to which the Lugar survey is still influencing perceptions on WMD risks is probably minor. In 2021, there was only one citation of the survey; in 2022, there were none; and in the first half of 2023, there were two citations. This suggests that the time might be right for a new study, carefully conceived and implemented, that addresses WMD risks in the current era and near future.

In contemplating the possibility of a new study, we have considered how to avoid repeating the misuses of the Lugar survey. We offer the following lessons and recommendations.

- **Responsibility of the survey team.** Perhaps the most important lesson for the survey team is that even an admittedly nonscientific survey will be treated by many as having scientifically valid results. Caveats about methodological limitations are readily cast aside, and the simplest forms of results are propagated in the literature. Survey teams must be fully cognizant of the public's desire for easy answers to complicated questions, of the media's desire for sensation, and of experts' tendencies to cherry-pick results that favor their preconceived views. Our recommendation is that future surveys be conducted with the goal of developing analytically rigorous results. This requires that they be conducted consistent with the best elicitation and analysis practices, many of which are described in the following section of this paper, in combination with scrupulous (even at the risk of being annoying) articulation and repetition of survey limitations. Future surveys should be designed so that it is not easy to misrepresent their results.
- **Responsibility of survey participants.** The Lugar survey participants came from many disciplines within the overall field of international security. But it is eminently clear that no single participant had expertise across the broad spectrum of questions posed in the Lugar survey. At the same time, almost all participants answered almost all questions. So we conclude that many of the responses to many of the questions are capturing nonexpert views. Much of the responsibility lies with the survey organizers who put participants in the position of either responding to questions in areas outside their expertise or appearing uncooperative. But some of the responsibility falls on the participants; they should have refused to answer questions outside the scope of their expertise.
- **Responsibility of citing authors.** Authors who uncritically cite the Lugar survey also share responsibility for misuse of the survey results. Such authors blatantly ignore Lugar's clear caveat about it not being a scientific survey, thereby

> misrepresenting the validity of survey's findings. This misuse of survey results perpetuates a dangerous, alarmist, and shortsighted perspective of WMD risks. Authors who cite the Lugar survey should acknowledge the limitations of its methodology and emphasize the variability of responses. More generally, they should invoke at least a minimal level of healthy skepticism about the survey results, as they should with any analyses they cite. It is not clear whether authors who have cited the survey were incapable of recognizing the survey's limitations and flaws or whether they ignored them when the survey's results supported their views. Neither possibility is comforting.

It is worth emphasizing that the Lugar survey team, the responding experts, and citing authors are all subject to personal biases. Undetected or unchecked biases can falsely influence any reader's agreement or disagreement with the results and can contribute to a false credibility for the Lugar survey and its report.

In summary, Senator Lugar intended to create a survey that raised awareness, both within the general public and the policymaker community, on the risks associated with WMD proliferation and use. The survey report identifies the purpose of the results as being "useful in helping to define the parameters of the risks that we face, assessing the current state of non-proliferation and counter-proliferation efforts, and identifying issues of concern that require more attention."[60] Although Lugar's intentions were carefully stated, the common uncritical and oversimplified representations of his findings encourage misuse of the survey's results. When considering similar future efforts, the survey team, participants, and citing authors alike should be much more diligent in representing the limitations of the results.

Formal Elicitation of Expert Knowledge Topics

While it would be interesting to ask the Lugar survey questions to the same experts today to see whether or how much their answers would change, the purpose of a follow-up survey would probably change and so would the experts, precluding such a direct comparison. Instead, a rigorous, defensible study using formal survey instrument[61] and elicitation methods should be planned and implemented. Such an effort provides a defensible foundation of experts' knowledge that can be updated with observed data, information, and evolving knowledge.

This section briefly presents formal elicitation principles and methods to demonstrate how knowledge can be acquired with bias-minimizing methods and how this information can be subsequently analyzed. The purpose of contemplating a next knowledge acquisition study (e.g., survey instrument) is to obtain the best-quality knowledge from properly chosen experts using a verified[62] and defensible[63] methodology and to then document that knowledge in a knowledge base.

Many of the elements and topics of a formal elicitation study presented in this section compare and contrast preparations for a new study with the Lugar survey.

[60] Lugar, *Lugar Survey*, 4.

[61] The term *survey instrument* describes the questionnaire, response mode, and corresponding instructions for the expert, whether for a survey or an interview in an elicitation.

[62] Methodology originally commissioned by the Nuclear Regulatory Commission for enhanced probability risk assessment for reactors and resulting in the book Meyer and Booker, *Eliciting and Analyzing Expert Judgment*.

[63] The 1999 R&D 100 Award–winning PREDICT methodology, invented by Thomas R. Bement, Jane M. Booker, William J. Kerscher III, and Mary A. Meyer, validated the elicitation methods.

Carefully Apply Terminology

Words like *judgment* or *opinion* should be avoided because they connote uninformed information, "the man on the street" responses, and wild guesses. Instead, words like *expertise*, *knowledge*, *informed judgment/opinion*, and *experience* convey and recognize the knowledge, experience, and expertise inherent in experts' responses and inform the expert of the quality of information being elicited.

Accurate word choice throughout the survey instrument is important to convey and communicate meaning between the experts and the survey administrators or analysts. For example, the Lugar survey questions and text used the words *opinion* and *judged* numerous times. To recognize the qualifications of these experts and to motivate them to provide thoughtful, informed responses, other words, such as *knowledge* and *expertise*, would have been better choices.

Every community of experts has terminology that is understood by its members, called the "community of practice."[64] This terminology should be used in the instructions for the survey, the survey itself (called the survey instrument), and in the documentation and reporting of the results.

If any terminology is ambiguous, broadly defined, or not widely used, specific definitions should be provided. Examples of words and phrases in the Lugar survey that would have benefited from specific definitions include *vulnerability*, *weapon of mass destruction*, *nonproliferation*, *threat*, *risk*, *likely*, *victory*, *disarmament*, *nuclear explosion*, *terrorist*, *possibility*, *national security*, and *a government*. Improper use of terms or poorly worded language could cause experts to question how their responses will be used by those who are not perceived as knowledgeable in their area(s).

How can one ensure that the proper definitions and terms are used to avoid miscommunication and lend credibility to the survey? The answer is to use an advisor expert. An advisor expert is a friendly, cooperative expert who is a valuable member of the elicitation team. Duties of this role include the following:

- providing an entrée into the community of experts
- identifying the experts
- providing definitions, jargon, and terminology
- aiding in formulating the questions and in choosing the appropriate response modes
- aiding in motivating the experts to participate
- aiding in interpreting the responses
- aiding in follow-up interactions with experts
- being the first to test (i.e., pilot test) the drafted questions

Multiple advisor experts may be necessary to represent a variety of expertise areas.

It is not known whether such a person was consulted during design of the Lugar survey. However, Senator Lugar himself could have fulfilled that role, representing the lawmaker community. As a recognizable advocate in the nonproliferation community, he provided credibility to the survey and also motivation for experts to participate.

Identify Experts

The first question asked is: What is an expert? An expert is a person recognized by their peers as having knowledge and experience in their field.

According to this definition, an advisor expert is suitable to assist with identifying other experts. Self-identified experts are less desirable. Choosing experts only known to the designers and administrators of the survey instrument, as appears to

[64] *Community of practice* is a term from Holland and Quinn, *Cultural Models in Language and Thought*. It refers to people's customs, artifacts, oral traditions, what they know to act as they do, and how they interpret their experience in a distinctive way.

have been done with the Lugar survey, results in a biased set or a bad sampling of the entire community of expertise. Results from a biased sample are not valid.

As evident in the Lugar survey, experts from multiple subject areas are needed to cover all aspects of the complex nonproliferation problem under study. It is unlikely that any single expert has the experience, knowledge, and information sufficient to cover the broad range of subjects encompassed by all the Lugar survey questions. For future studies, either the scope of the study subject should be narrowed or different experts should be asked about only the subset of content that aligns with their particular area(s) of expertise.

Construct a Representative Sample of Experts

A statistically valid sample is a selection of experts from the entire population of all experts such that the chosen set of experts represents the various characteristics of the whole population. In the unlikely event that there are a large number of experts, a scientific sample using sampling techniques is recommended.[65] For small populations of experts, selecting an unbiased large majority, including striving to persuade all (a census) to participate, suffices.

Striving for a representative sample differs from selecting a *convenience sample*—for example, a sample of experts known to only one person may or may not include the most esteemed experts within the given field. For any population size, one would want to select as many experts as possible.

For the Lugar survey, it is not known how many experts existed in 2005, so it is not possible to judge whether the eighty-five respondents were a reasonable or representative percentage of that population. In many fields of expertise, eighty-five of one hundred thirty-two would be a large percentage of the targeted population. Those in the Lugar set of one hundred thirty-two were chosen for their diverse viewpoints, different countries of residence, and varied careers, making the set likely to have been a representative sample of the targeted population of "elites."

Motivate Chosen Experts to Participate

The general rule is to motivate the selected experts to participate during the first contact and to recontact those who fail to initially respond. While the Lugar survey report does not state what efforts were made to recontact and encourage participation from the experts who failed to initially respond, we learned that efforts were made and they resulted in the survey's good response rate. The late Senator Lugar personally contacted some experts. Follow-up contacts provide an opportunity not only to motivate the experts but also to clarify any misunderstandings and address concerns.

Properly motivated participation in the first place reduces the number of follow-up contacts and nonresponses. Everyone (expert or not) likes to be flattered and made to feel important. This is not a false flattery because the knowledge and information these experts have is really important to capture. The advisor expert can aid in crafting wording that motivates participation and then in recontacting those who did not initially respond.

Despite the best efforts to motivate them, some experts will refuse to participate. Starting with a larger-than-necessary list of experts minimizes the impact of the loss of knowledge from nonrespondents. Care must be taken to ensure that all who refuse are not all of one type or group of experts; such a loss produces a biased, nonrepresentative sample.

[65] Schaeffer et al., *Elementary Survey Sampling.*

Understand Common Biases in Expert Knowledge

Unfortunately, all humans think and act through the filters of personal biases. Experts are no exception. Bias is defined as a slanting, adjusting, or filtering of the expert's thinking and original knowledge owing to their perspectives (motivations) and thought processes (cognition).

Cognitive and behavioral biases that particularly emerge in a mail-in survey or interview include:[66]

- *Anchoring bias*, a cognitive bias that results from a failure to adjust from the expert's initial or first impression despite alternative or new evidence. A strong personal agenda causes the expert to remain anchored to their view regardless of and despite new, contrary knowledge and/or information.
- *Wishful thinking*, a motivational bias that results in an expert's tendency to allow their hopes to influence their answers, methods, desires, decisions, results, and conclusions.
- *Availability bias*, a cognitive bias that results from how easily an expert can retrieve particular events from memory. In particular, this bias affects how accurately probabilities or percentages can be estimated. Because memory by its nature is selective, recent experiences tend to dominate the process of estimation.
- *Underestimation of uncertainty*, a cognitive bias that is one of the most commonly occurring biases in analysis, assessment, and quantification. Humans think the world functions more precisely than it does, based on their experiences, expectations, and memories.
- *Training bias*, a motivational bias that results in the data gatherer's (elicitor's) and/or analyst's tendency to misinterpret data or information from an expert for their own purposes. For example, an analyst can group qualitative answers to a question in such a manner as to support a particular outcome, distorting the original intent of the experts.
- *Impression management*, a motivational bias that results from social pressure. It occurs when the expert responds to the reactions of those who are not physically present. For example, the expert answers the survey questions in a way that maximizes approbation either from society in the abstract or from the administrator of the survey in particular (e.g., Senator Lugar).
- *Inconsistency*, a cognitive bias that results in the inability to maintain the same problem-solving, heuristic, definitions, or assumptions throughout the duration of the survey/interview because of the human mind's limited information-processing capability.

Great care must be taken in the design, implementation, and analysis of the elicitation to understand, monitor, and minimize cognitive and motivational biases. Apparently, neither bias detection nor bias minimization was undertaken during design of the Lugar survey. Yet, techniques are readily available and can reasonably be implemented. Without the employment of such methods and techniques, the information elicited will not be a true representation of the experts' thinking or knowledge.

Choose an Appropriate Communication Mode

Many forms of communication are available—for example, telephone calls, videoconferencing, face-to-face interviews, and mail-in surveys. Of these, face-to-face interviews provide the best opportunity to obtain the best-quality responses through the use of bias minimization elicitation techniques, whereas mail-in surveys provide the least opportunity.

[66] Meyer and Booker, *Eliciting and Analyzing Expert Judgment*, chap. 3.

The Lugar survey report does not specifically state how the survey was administered, but it is implied that the survey was mailed. A mail-in survey is the most difficult communication mode for monitoring and minimizing biases that shift experts' answers away from their true knowledge. Examples of problem areas include inaccurate or ambiguous terminology and poorly formulated questions.

Provide Instructions and a Cover Letter to Experts

A cover letter was sent to the Lugar survey experts. Any cover letter should include instructions for the experts, including a schedule for returning responses. The following information should be provided to experts as part of the survey instrument:

- The purpose of the survey
- Why eliciting expert knowledge is important
- What people or organizations are involved in administering the survey
- Why the expert being addressed was chosen
- Motivation for experts to participate
- An estimate of the time required to complete the survey
- How and when to return survey responses
- How the information provided will be used
- Whether the expert's name will be used (either collectively or individually)
- Whether answers and/or information provided will be listed collectively or individually and whether the expert's name will be attached to anything
- When the experts will see what was done with their information
- Whether experts be given an opportunity to edit or revise anything
- Contact information for someone who can answer any questions the expert has
- What assumptions the expert should make about the subject matter or particular questions
- Relevant definitions of terms
- Relevant background material for experts to use as they see fit
- An indication of whether experts can confer with others
- Assurance that the experts' wishes and the survey administrators' promises will be honored

Of the above items, the Lugar survey report mentions that the experts agreed to their names being associated with the compiled results—a good and necessary step taken by the survey designers and administrators.

Control the Length of the Survey Instrument

Regardless of the communication mode, a survey instrument should be long enough to capture the necessary information in an unbiased manner but short enough to not fatigue the experts responding to it.

The twenty-one-question Lugar survey is a reasonable length. However, the broad nature of the questions does not tap deeply into the experts' valuable knowledge and does not capture important information from them. Some experts may not have responded because the questions were so general. The danger is that the expert may interpret a survey's generality as naivete of its designers or administrators.

To reduce the number of questions, different questions can be posed to experts depending on their area(s) of specialty. To gather more detail on the subject of a question, sectioning the survey instrument is a solution for tapping more deeply into the experts' knowledge without fatiguing them. An example of a sectioning structure would be to ask the broad question first and then instruct the expert to skip subsequent detailed questions if they are not comfortable delving deeper.

Design Questions Using the Experts' Terminology

Surveys should not include leading, agenda-driven, slanted, ambiguous, unclear, badly worded, or insulting questions. Telemarketing push polls are a prime example of these kinds of questions. To avoid those pitfalls, however, surveys often ask only broad, sweeping general questions, as was done in the Lugar survey. These broad questions do not capture important knowledge and thinking of the experts. Questions must be structured to minimize bias and elicit specific knowledge using formal techniques, and they must be cast in terms familiar to the experts.

While good questioning starts at the general level to establish the subject, additional questioning is necessary to drill down to the specifics, uncovering the limits of knowledge and the thought processes of the experts. Studies[67] have shown that the best knowledge (especially for complicated and poorly understood problems) is obtained by decomposing the problem into specific parts and details. No such attempt was made with the Lugar survey's twenty-one general questions.

For example, consider question 8 in the Lugar survey. Several questions should have followed this one, asking about specific mechanisms for a terrorist to acquire nuclear weapons and specific problems or issues with manufacturing them. After getting the expert to think through the detailed questions, one would then ask the expert to return to and potentially revise their initial assessment to the general question 8. In addition, the wording in question 8 suggests only two options: "acquisition" *or* "manufacture." What about a combination of those? The same is true of question 6. Could there be a terrorist–government hybrid answer? This limitation on response choices is discussed further in the subsection on response modes.

In question 19, the union of two different entities, the United States and the international community, presents a difficulty for the experts because priorities for the United States may not be identical to those of the international community, nor would their highest priorities necessarily be the same. This question could have been split into two parts: one for the United States and one for the world. Analyses would then determine whether the experts' responses were the same for both.

Questions 19 and 20 exemplify one-sided or biased questions. Question 19 does not accommodate any expert who views proliferation positively or nonproliferation as not so important. The same is true for question 20, which does not permit risks to be underrated. Even though the wording of these questions may be designed for the survey goals, the phrasing of the questions should remain as neutral as possible to minimize confusion and bias.

Another example of a potential question-phrasing problem is found in the stated result of question 10: "Expectations . . . were widely dispersed." Looking at this question, words used like *numerous* and *major* are not well defined. To obtain more precise (less widely dispersed) answers, more precisely articulated questions must be asked.

There appears to have been miscommunication in question 17 as well. The question begins, "If you answered [to question 16] too much or not enough spending." Following these instructions, twenty-six experts should not have answered this question, yet all the experts provided an answer, ignoring the instruction.

Questions 18 and 19 contain one-sided phrasing, cutting off the options of the other side. This

[67] Meyer and Booker, *Eliciting and Analyzing Expert Judgment.*

induces a bias. Question 18 asks the experts to think in terms of “encouraging developments” without following up with discouraging ones. Question 19 defines “non-proliferation” as the “goal.” Not everyone agrees with that; some experts emphasize benefits from proliferation. These phrasings and the choices listed in question 18 reveal the intent of the Lugar survey, which is not consistent with acquisition of experts’ knowledge. The questions should have been pilot-tested to uncover any bias that would slant the experts’ responses in the direction of the Lugar survey’s purpose.

The phrasing of question 21 is a curious mixture of specific and general. The question specifies “during the last year,” but it does not specify to whom or for what purpose their recommendation applies. Should experts cite material for the general public, their peers, their government, Senator Lugar, or the US government? If experts were asked to provide reasoning and sources used in answering the other questions, then question 21 would have already been answered.

Apparently, there was some difficulty with some responses, as noted on page 4 of the survey report: “In a small number of cases (fewer than 10) specific answers to individual questions were not included in overall calculations due to discrepancies or miscommunications.” These responses should not have been omitted. Instead, the experts should have been contacted and the issues resolved.

Construct a Response Mode

The form or forms of the requested responses should be clear, easy to use, convenient, and customary for the expert. The choice of specifying how the experts respond to each question—called the *response mode*—is as important as the choices of words in question phrasing. Response-mode choices should not be made primarily for the convenience of the survey designer, analyst, administrator, or promoter. Question phrasing can restrict the choices of response modes. As noted above, for example, the “or” question phrasing in the Lugar survey’s question 8 limits the expert’s choices when responding.

The Lugar survey used a variety of response modes, which is generally good; however, they fell short of capturing the experts’ knowledge, insights, and understanding.

As previously noted, the Lugar survey used three major groups of questions corresponding to three types of response modes:

(1) Questions asking for unrestricted (open) numerical responses (questions 1–14 and 17)

(2) Multiple-choice questions (questions 15, 16, and 18)

(3) Open questions requesting written (qualitative) responses (questions 19–21)

Questions 1, 2, and 3 were clearly worded with an open response mode, asking the experts to plainly specify “how many.” Questions 4, 5, 9, 10, 11, 12, 13, and 14 asked the experts to provide a “probability.” In questions 4, 5, 9, and 10, “probability” was to be expressed as a “percentage.” While this important clarification for the response form was provided for these four questions, this instruction was dropped for questions 11–14.

Probability and percentage are technically not the same. Most experts do not estimate probabilities accurately. Unless probability is a fundamental topic in the experts’ community of practice, it is best avoided as a response mode. Alternative choices include propensity, likelihood, proportion, and percentage. Unfortunately, the Lugar survey conflated percentage with probability. If a simple percentage (from 0% to 100%) is desired, then that is a reasonable response-mode choice. Even with something as commonly used as percentage, it is helpful to clearly define what the percentages mean (e.g., 0% means the event never happens, and 100% means that it happens without a doubt).

As emphasized in the analysis section of this paper, uncertainties exist in human thinking, and experts should be asked to describe their uncertainties in their responses. A straightforward response mode for capturing uncertainty is to ask experts to provide their "best" (or "most likely" or "middle") estimate first and then to provide both "low" and "high" estimates, representing their uncertainty about their estimate. Asking experts to state their best estimates first helps to capture their initial thinking.

For example, Lugar survey question 1 would have three parts: the first would ask the experts for their best estimate of the number of nations, the second would ask them to estimate the minimum number, and the third would ask them to estimate the maximum number. Before asking for these three responses, the survey should alert the experts that these three responses will be requested and explain why. If experts are told that single and range estimates are going to elicited, they may choose to begin with ranges, which is fine because that is their choice.

The multiple-choice response modes in the Lugar survey questions 7 and 18 restricted experts to only one-choice answers (unless they specified "other" as multiple responses). It is best to ask experts to select as many answers as they deem appropriate and then to rank them in order of importance (i.e., priority). Tied ranks should be permitted. An enumeration across experts for each choice can be determined by using the experts' ranks to form weights.[68]

Lugar survey questions 15 and 16 imply an ordinal response mode: "too much," "too little," or "just right." While those crisp and distinctive choices were fine for Goldilocks, when seeking expertise, it is preferable to give the experts more choices. A five-point scale (far too little, too little, just right, too much, far too much) or an expanded seven-point scale increases response options. Better still is to offer experts a continuous scale, a line with the end and middle points labeled. Experts mark their answer anywhere along the line with an *x*, and then they mark the range of values with a line segment around it. The continuous scale avoids the discrete structure of response choices, and it provides flexibility for the expert's response.

A continuous scale could have been used in question 17; however, leaving the response mode unspecified (open) is also a reasonable option for that question.

Whatever response choices, whether quantitative or qualitative, are offered to the expert, they should be familiar to the expert in their community of practice. For example, do not provide a number line from 0 to 1 for a question when experts think in terms of answers like 10^{-5} or 10^{5}. For those situations, a log scale (which can be displayed as a linear scale of orders of magnitude) would be appropriate. Requesting percentages, as done in the Lugar survey, precludes experts from responding in any nonlinear or orders-of-magnitude scales, which may correspond to their thinking. As discussed, the advisor expert can assist with determining and constructing the appropriate response modes.

Questions 18, 19, 20, and 21 contain qualitative response modes, as opposed to quantitative modes. Of these, question 18 has the most desirable mode, offering several examples of choice and including an open-ended "other?" choice that can be specified by the expert. Questions 19–21 are open-ended, allowing the expert to provide qualitative responses. Open-ended questions give the expert freedom to respond; however, that freedom can also lead the

[68] For example, suppose expert A only selects choice 3. Expert B selects choices 1, 3, and 4, ranking them as middle (second-most important), top (most important), and bottom (least important), respectively. The weight for expert A's choice is 1.0 for 3 and 0 for the other choices. To be fair, the weights for expert B's choices are constrained to sum to 1.0; therefore, the top ranked choice (3) has a weight of 3/6 = 0.5, the middle ranked (1) has a weight of 2/6 = 0.3333, and the bottom ranked (4) has a weight of 1/6 = 0.1667. Tallying across both experts produces a total of 1.5 for choice 3, 0.3333 for choice 1, 0.1667 for choice 4, and 0 for all other choices.

expert to wander off into a different subject area. Careful response-mode planning (usually with the advisor expert) is required to achieve the balance of properly guiding the expert to respond in their own way without biasing or leading them.

The difficulty with open-ended (essay) questions is that the analyst needs to have information about the clarification, definitions, assumptions, and problem-solving processes used by the expert in order to analyze the qualitative responses. In a face-to-face (including video) interview, such issues can be determined through detailed questioning. In a pilot-tested and well-designed set of questions, this knowledge can also be elicited.

The following are examples of additional questions that command consideration:

- Are the experts all answering the same question? Subtle differences, assumptions, and problem-solving processes can cause an expert to answer a different question than the one being asked.
- Are any assumptions or commonly known facts driving their answers? These are the basis for eliciting problem-solving processes in the experts' thinking.
- Are they providing the accepted "party line" or providing their honest answers? Are experts providing their own answers or quoting others? This is a common bias requiring detection and minimization.
- Are the provided answers legible and understandable? Clarification of the experts' responses requires probing, which can also be done after the survey or interview.

Without addressing the above questions, the valuable, unbiased flow of knowledge from the experts' thinking through to their responses is lost.

Responses, such as the goals elicited in question 19 and the risks elicited in question 20, are difficult to categorize into distinctive (crisp) sets for analysis. For question 19, the analyst(s) obviously selected a categorization that corresponded to the Nunn-Lugar programs and objectives, as explained on page 30: "More than a quarter of respondents (27 of 85) either listed by name the Nunn-Lugar Cooperative Threat Reduction Program . . . or listed as the goal a particular Nunn-Lugar objective." While such a category favors the Nunn-Lugar program, it is a composite of separate goals, making for a very broadly defined category or bin. The other goals provided by experts may not have been so broadly defined. Such a categorization mixes levels of detail for the bins: the top goal is a broad collection of Nunn-Lugar goals and the other goals are individual goals. This is an example of mixing granularities. Requesting a consistent level of detail not only aids the analyst but also avoids confusing the expert.

The boundaries of the chosen categories for question 19 appear to be vaguely defined. In other words, the analyst may not be able to precisely determine whether a given response fits exclusively into the Nunn-Lugar category or into another category. Such difficulty of classification into categories makes this a candidate for fuzzy set theory application.[69] In fuzzy set theory, an element (a response) can have partial membership in more than one set (category). For question 19, a provided response may have partial relevance to one of the Nunn-Lugar goals and also partial relevance to another goal. The Lugar survey report does not say whether any expert provided more than one goal in response to this question. Because question 19 is *the* major question of the survey, it is unfortunate that its results were not more thoroughly discussed and analyzed.

Looking at question 20, it is unknown whether the experts provided valuable information about why they responded in the ways they did. What was the reasoning for their responses? Did anyone provide a new, different, or unique response? What was learned

[69] Zadeh, "Fuzzy Sets."

from these responses, individually or collectively, that was not previously known? Were any difficulties with terminology encountered in the responses? Did any experts provide more than one answer? As with question 19, important expert knowledge was not given due attention and may not be documented for future reference and understanding.

Ask for Experts' Thinking, Reasoning, and Problem-Solving

It is not known whether any attempt was made to gather information from the experts about their thinking in answering each question in the Lugar survey. In a face-to-face interview situation, it is easy to continuously ask the expert what they are thinking, what assumptions they are making, what resources or experience they are relying on, what theory they are using, and what their reasoning and problem-solving processes are as they consider and respond to each question. Gathering that kind of detail is difficult in a mail-in survey, but some questioning along those lines can be included as additional questions.

Gathering information about the expert's thinking, reasoning, and problem-solving helps to minimize different kinds of biases, ensures that the expert is answering the correct question, provides traceability and a memory trail for future reference, makes it easier for the expert to update their answer later, assists with resolving differences in experts' answers, and ensures a thoughtful, good-quality response.

Furthermore, it has been shown that experts' responses are not well correlated (associated) based on common background (e.g., college) but by the ways in which they solve problems.[70] When experts disagree, the reasons for this disagreement often emerge from their using different reasoning when answering the question. Because of differing cognitive processing, experts are actually answering slightly different questions, resulting in their different answers.

Be Aware of Difficult and Sensitive Questions

Decision-makers and technical professionals, especially in the risk, reliability, and safety communities, often ask how they can know what they do not know (the unknown). This is usually in response to experts experiencing a rare, unanticipated, previously unknown event, called a black swan.[71] The attacks on September 11, 2001, are an example. However, 20/20 hindsight often reveals that someone, somewhere *did* think of the improbable beforehand but was either not heard or was afraid to admit to thinking outside the norm, the party line, or the politically correct. While it is easy for experts to evade difficult questions by saying "I don't know that," usually they reveal what they do know through continued conversation. Experts can be encouraged to think of the impossible and venture beyond their comfort zones as long as that information is handled with appropriate care. These kinds of imaginative discussions can produce gems of knowledge, which can avoid the unexpected.

Unfortunately, no such difficult questions were asked in the Lugar survey; yet it is not difficult to imagine that such questions exist in the nonproliferation community. Without difficult questions, the only answers that will be collected are mostly the usual, expected ones. Insights, new knowledge, and thoughts on the unknowns are *not* captured by such comfortable questioning. To get information about the unknowns, difficult questions must be asked.

Great skill is required to ask difficult and probing questions, and the best format is face-to-face. The advisor expert can help determine appropriate question phrasing and how to carefully and

[70] Booker, Meyer, and Martz, "Sources of Correlation of Expert Opinion."

[71] Taleb, *Black Swan*.

comfortably probe the expert for answers. Assurance must be provided because often the expert must expand their thinking beyond the established norm or accepted theory—even into the realm of the politically incorrect and things no one wants to consider. Experts must be carefully nudged, flattered, and encouraged to speculate, "think outside the box," broaden their thinking, or be creative while being promised that what they say will not come back to haunt them later (and that it may, in fact, come back to confirm them). One way of providing assurance is to protect anonymity. Another way is to elicit the uncertainty or conditions and caveats that the expert wants to be attached to their answer. We are all familiar with how the detective must think like a serial killer to stop one. The same is true of a terrorist. Asking this of the expert is asking a lot, but who else is qualified to provide an answer, even a highly uncertain one?

Perform a Pilot Test

Once the instructions and cover letter are ready, the terminology and definitions are established, and the questions and response modes are finalized, the survey instrument is ready for its first complete test—the pilot test. The advisor expert is perfect for this task. The advisor can identify bad question phrasing, poorly defined terms or words, clumsy or inadequate respond modes, missing information or assumptions, and potential sources of bias.

It is not known whether the Lugar survey was pilot-tested before being sent to the experts. If it was not, perhaps such a test could have prevented some of the confusion regarding some of the responses that were omitted.[72]

Having addressed these issues, the survey instrument is ready to be sent to the experts or used in an interview session with them. Those interviews can be audio only, video, or physical face-to-face sessions and must be scheduled at the convenience of the expert. Additional information on conducting the elicitation guidance is provided in Meyer and Booker, chapter 10.[73]

Provide Feedback to Experts

Experts should be shown exactly what was done with the information they provided—they should be given feedback. While this step occurs after the results are gathered and analyses are done, it should be implemented **before** any public dissemination of the survey and its results.

Feedback gives the experts and the analyst the chance to review, revise, edit, or clarify. Feedback demonstrates to the experts that promises were kept, that proper care was taken with their responses, and that trust was earned. It is not known what, if any, feedback was given to the experts before the release of the Lugar survey report.

Document Everything

All the preparation, execution, questions with responses, and knowledge gathered from the survey should be documented, in real time, in a knowledge base or repository[74] for future reference and use. Material for documentation includes everything from the original idea of why such a project was necessary to the final report. All the choices and decisions made should be recorded, such as the choice of survey instrument, the selection of experts, and the content of the questions. Documentation is best done while the work is ongoing. It is more difficult to do after the work is completed, when schedules and funds are ending. Invariably, at

[72] "In a small number of cases (fewer than 10), specific answers to individual questions were not included in overall calculations due to discrepancies or miscommunications." Lugar, *Lugar Survey*, 4.

[73] Meyer and Booker, *Eliciting and Analyzing Expert Judgment.*

[74] A knowledge base can be as simple as a file folder in a cabinet containing all the notes, documents, data, and analyses or as complicated as an interactive, user-friendly online database such as one of the commercially available software options.

some point in the future, this documentation will be invaluable because someone will ask probing questions, wondering, for example, why something was done the way it was done, what something means, or what someone was thinking back then.

It is not known what additional documentation was prepared outside of the final Lugar survey report. The original responses were, in fact, archived,[75] but they were not made available to us.

Design the Analysis with the Elicitation

Expert knowledge, whether qualitative or quantitative, elicited from experts is like any other kind of data and, therefore, can be analyzed.[76] However, the treatment, interpretation, analysis, and use of the data gathered from the experts' responses require as much attention to proper use of formal analysis techniques as does the elicitation itself.

These two major efforts, elicitation and analysis, are inexorably linked. How the data are handled and analyzed depends on the information content inherent in them, which, in turn, comes from the information content in the questions processed through the experts' brains. Said another way, to get the desired data out, one needs to ask the right questions of the right experts.

As already noted, important in this link between elicitation and analysis is the response-mode choice. For example, the analyst cannot arbitrarily transform qualitative answers, such as those in the Lugar survey question 20, into numbers for analyses. If numerical responses are desired for a question, they must be elicited using a numerical response mode. Only the expert can quantify their qualitative knowledge.

Terms used in common parlance may or may not be appropriate for the community of practice of the chosen experts. Common examples include *probability*, *risk*, and *uncertainty*.

Unfortunately, the term *uncertainty*[77] is not mentioned in discussion of the questions or responses in the Lugar survey report. Yet on page 4, in the fourth paragraph, the author of the Lugar survey report states an intent to "discover consistencies and divergences in attitudes." And for question 12, on page 21 of the survey report, the author concludes that "the range of responses broadened." These both are statements about uncertainty; yet, it appears that nothing was done to address it. All knowledge, information, and data have some degree of uncertainty attached to them. It is important to identify the kind of uncertainty involved, to elicit the degree of uncertainty, or to state that an uncertainty is assumed to be negligible (if that is a reasonable assumption to make).

The last thing an analyst wants is to receive responses from experts that violate mathematical principles (e.g., a probability > 1.0). The last thing an expert should experience is being asked to provide an answer in a form they do not understand. The steps outlined in these subsections are designed to avoid such occurrences.

Risk is a difficult term to use unless it is part of the experts' community of practice. Even then, a subtle definition reminder should be provided. For example, risk is a compound quantity: the *likelihood* of an event and its adverse *consequences*. Unfortunately, the Lugar survey seems to equate risk only with chance or likelihood in the discussion and in the titles of the questions.

As noted in this paper's section on uses and abuses of the Lugar survey, displaying the results for feedback to the experts, for publication, for the public, or for stakeholders must also be carefully

[75] Conversation with Dan Diller, director of Policy at the Lugar Center, on August 11, 2022.

[76] Booker and Meyer, "Using Expert Judgment as Data."

[77] Uncertainty is that which is not known precisely. Uncertainty includes unknowns from a variety of sources, including lack of knowledge; poorly understood physical relationships, theories, or behaviors; random variability or noise level; nonspecificity; and lack of data.

considered to ensure that the desired outcome is conveyed and misunderstandings are avoided. The experts will want to see how their inputs were used and displayed to ensure the practices are in accordance with their community of practice.

The choices for display of the Lugar survey results appear to be reasonable, assuming binned charts and pie charts are well accepted practices within the diverse nonproliferation communities of the chosen experts. The charts are easy to read, well marked, and discussed. Displaying one question and its results per page (or more) is convenient.

However, aggregated displays and final results (e.g., averages) come at the cost of losing valuable details in the original experts' responses. For example, the binned charts in the Lugar survey report displayed separate 0% and 100% bins but did not separate out a 50% bin.

Responses to the Lugar survey question 17 are shown with collapsed bins of 25% intervals and with all the answers 100% and up in one bin. This is a loss of detailed information because that bin consequently has the largest frequency (becoming the most likely answer). How many experts supplied answers far beyond 100%? What was the maximum? Did large percentages correspond to fewer and fewer respondents? Potentially important information was lost with this collapse. In addition, the average calculation for this question was not supplied, which would have somewhat indicated the distribution of responses in that large bin.

For question 18, a binned chart could have been used to display the answers, offering visual understanding and following the established precedent for other response displays.

When reporting results, *all* the written responses should be included unless there is a good reason not to do so (e.g., the expert requests omission). Some of the answers are listed for Lugar survey questions 19, 20, and 21. Every participant who reads this report needs to see that their input was valued as much as another's, even if their answer was the only one of its kind. When interpreting the results, the analyst should always pay attention to the "lone ranger" who provides an answer that seems to go against the "norm"—that original thought may prove to be the most important one. While such one-of-a-kind answers present a challenge for the analyst, they are a part of the total state of knowledge about the subject at that time. They should all be documented. The purpose of any elicitation is to capture the state of knowledge at that time.

Final Thoughts on Future Elicitations

The goal of the Lugar survey was to "contribute to the discussion inside and outside of governments about how we can strengthen non-proliferation efforts, improve safeguards around existing weapons and materials, bolster intelligence gathering and interdiction capabilities, and expand international cooperation in dealing with a threat that should deeply concern all governments and peoples."[78] Questions and experts were chosen to emulate a hearing and to contribute information of value to programs and efforts on nonproliferation.

At the same time, the survey was not intended to be "scientifically" rigorous. Thus, its results should not be taken at face value, but instead as a starting point for further discussion and analysis. Accordingly, this paper draws lessons from the design and execution of the Lugar survey for the planning of future studies. The most important lessons include the following:

(1) Select experts representing the broadest, most general population of all perspectives and understandings of the issues.

(2) Segregate experts' questions according to their areas of expertise (e.g., do not ask chemical weapon experts about nuclear risks).

[78] Lugar, *Lugar Survey*, 1.

(3) Elicit experts' reasoning and problem-solving for each question.

(4) Design questions and response modes to include accompanying uncertainty.

(5) Incorporate bias minimization methods in question formulation and elicitation implementation.

(6) Analyze results using appropriate mathematical methods.

(7) Provide a repository of data accessible for further research by others.

(8) Anticipate misuse of the results by the academic and policy communities—emphasize all caveats.

It may seem that all our recommendations relating to the elicitation and analysis of expert knowledge are too numerous, too complicated, too impractical, or too difficult (or all of these) to completely implement. However, the Pareto principle applies—a little effort to accomplish as many of these suggestions as possible will go a long way in getting better-quality and more useful knowledge from experts.

Therefore, the most important directive for any knowledge acquisition study is to make the effort to follow the suggested steps contained herein for the gains they can provide without unduly bankrupting or delaying the project. Any shortcomings should be reported so that users of the results can evaluate the quality of the knowledge presented.

In terms of evaluating risks associated with WMD, we recommend undertaking *several* independent studies. Absent collusion, it seems unlikely they would come up with the same answers, and whatever disagreements exist will inspire further thinking and evaluation. Each study should address uncertainty in the data and information, with subsequent studies being designed to reduce those uncertainties, where possible. Several independent studies will also reduce the risk of rote recitation of the results of a single study, as has too often been the case with the Lugar survey.

Finally, we have come to the conclusion that asking experts to predict the probability of WMD use in the next five- and ten-year periods is unlikely to lead to a working agreement that would influence policy. Humans are subject to motivational and cognitive biases and are not especially good at estimating probabilities. We suggest reformulating the focus of future surveys to address topics that are less likely to have these problems. For example, in the case of nuclear risks, the following questions could be posed:

- What do you think the most likely pathway to nuclear war is today, and would that pathway change in the next ten years?
- Are central and extended nuclear deterrence adequately robust? If not, what changes in force structure, strategy, and/or policy do you suggest?
- How should the United States deal with the emerging tripolar nuclear world, with China, Russia, and the United States having comparably large nuclear arsenals?
- How can the three-quarters-of-a-century tradition of nonuse of nuclear weapons be extended indefinitely?

Experts' perspectives on these questions are likely to be far more illuminating than their guesses about probabilities of future events.

Senator Lugar was correct in asserting that WMD risks "represent a threat that should deeply concern all governments and peoples." His statement remains valid to this day; however, little has been done to acquire the knowledge and data necessary for assessing these risks. Knowledge acquisition and analyses must be thoughtful and rigorous to effectively contribute to WMD risk assessment and management. This paper, we hope, provides useful guidance for those future studies whose results will provide a sound foundation for discussion, research, policy formation, and legislation.

Bibliography

Allen, Lindsay, Mellissa Fratello, Julie Gotham, Hao Huang, Elea Mihou, Jody Pollot, Pavan Yadav, and Carol Yamarino. "What Dangers Do We Face? A Preliminary Multihazard Risk Profile for New York State." Presentation at the Department of Urban and Regional Planning Graduate Workshop, University of Buffalo, May 7, 2007; MCEER-07-SP01. https://www.eng.buffalo.edu/mceer-reports/07/07-SP01.pdf.

Anderson, Victor, and James M. Thomas. "Quantifying Unfortunate Events for Strategic Planning and Resource Allocation." California Department of Public Health, Sacramento, CA, 2008. http://slideplayer.com/slide/7408810/.

Bialik, Carl. "Pondering the Chances of a Nuclear Attack." *Wall Street Journal*, July 7, 2005. https://www.wsj.com/articles/SB112059629605777656

Booker, Jane M., and Mary A. Meyer. "A Framework for Using Expert Judgment as Data." *Statistical Computing and Statistical Graphics Newsletter* 2, April 1991.

Booker, Jane M., Mary A. Meyer, and Harry F. Martz. "Sources of Correlation of Expert Opinion: A Pilot Study." In *Advances in Risk Analysis 5: Risk Assessment and Management*, edited by Lester B. Lave, 345–354. New York: Plenum Press, 1988.

Brownstein, Charles, John Baker, Peter Hull, Nicholas Minogue, George Murphy, and Phyllis Winston. *Report of the DHS Small Vessel Security Institute*. Arlington, VA: Homeland Security Institute, 2007. https://apps.dtic.mil/sti/pdfs/ADA480860.pdf.

Bunn, Matthew. "A Mathematical Model of the Risk of Nuclear Terrorism." *Annals of the American Academy of Political and Social Science* 607, no. 1 (2006): 103–120. https://doi.org/10.1177/0002716206290182.

———. *Securing the Bomb 2007*. Cambridge, MA: Harvard University, 2007. https://www.nti.org/wp-content/uploads/2007/09/securing-the-bomb-2007-fullreport.pdf.

Chapman, Sally. "New Report Paints Grim Picture for Future WMD Attacks." HSDL blog, June 22, 2005. https://www.hsdl.org/c/new-report-paints-grim-picture-for-future-wmd-attacks/.

Defense Science Board. *Reducing Vulnerabilities to Weapons of Mass Destruction*. Vol. I, main report. Washington, DC: Office of the Under Secretary of Defense, May 2007. https://dsb.cto.mil/reports/2000s/ADA471566.pdf.

Digges, Charles. "US Survey: Next Decade Holds a 70 Percent Chance of a Nuclear Terrorist Act." Bellona Foundation, June 23, 2005. https://bellona.org/news/nuclear-issues/nuclear-agreements/2005-06-us-survey-next-decade-holds-a-70-percent-chance-of-a-nuclear-terrorist-act.

"Experts Assess Likelihood of Nuclear, Biological Attacks," SitNews, June 26, 2005, http://www.sitnews.us/0605news/062605/062605_nuclear_bio.html.

Frankel, Michael J., James Scouras, and George W. Ullrich. *The Uncertain Consequences of Nuclear Weapons Use*. National Security Report. Laurel, MD: Johns Hopkins University Applied Physics Laboratory, 2015. https://www.jhuapl.edu/sites/default/files/2022-12/UncertainConsequencesofNuclearWeapons.pdf.

Gongol, Brian. "Structural Failures of a Mass Evacuation by Automobile." Gongol.com (personal blog), posted October 2, 2005, updated March 5, 2008. http://www.gongol.com/research/disasters/evacuationbyauto/.

Halstead, John. *Safeguarding the Future: Cause Area Report*. London: Founders Pledge, Updated December 2020. https://dkqj4hmn5mktp.cloudfront.net/Cause_Report_Safeguarding_the_Future_bf4f5328f7.pdf.

Holland, Dorothy, and Naomi Quinn, eds. *Cultural Models in Language and Thought*. Cambridge: Cambridge University Press, 1987.

Huemer, Michael. *The Problem of Political Authority: An Examination of the Right to Coerce and the Duty to Obey*. London: Palgrave Macmillan, 2013. https://doi.org/10.1057/9781137281661_1https://link.springer.com/chapter/10.1057/9781137281661_1.

Johnson, Maribeth, and Mark Litaker. "A SAS® Program to Calculate and Plot (1-a)100% Widths of Confidence Intervals for Binomial Proportions." In *Proceedings of the 1999 SouthEast SAS Users Group (SESUG) Conference*. http://analytics.ncsu.edu/sesug/1999/106.pdf. Originally from Jerrold H. Zar, *Biostatistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Kunreuther, Howard C., and Erwann O. Michel-Kerjan. "Evaluating the Effectiveness of Terrorism Risk Financing Solutions." NBER Working Paper Series, working paper 13359, National Bureau of Economic Research, Cambridge, MA, October 2007. https://doi.org/10.3386/w13359.

Lugar, Richard G. *The Lugar Survey on Proliferation Threats and Responses*. Washington, DC: US Senate, June 2005. https://irp.fas.org/threat/lugar_survey.pdf.

Meyer, Mary A., and Jane M. Booker. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. London: Academic Press, 1991; reprinted by Philadelphia: American Statistical Association-Society of Industrial and Applied Mathematics, 2001.

Michel-Kerjan, Erwann, and Debra K. Decker. "The Economics of Nuclear Energy Markets and the Future of International Security," Working Paper 2008-01-08, Wharton University of Pennsylvania, Philadelphia, January 2008. https://ciaotest.cc.columbia.edu/wps/isp/0002779/f_0002779_1947.pdf.

National Institute for Public Policy. *Minimum Deterrence: U.S. Nuclear Weapons and the Priority Threat Facing the United States*. Fairfax, VA: National Institute for Public Policy, September 2014. https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Litigation_Release/Litigation%20Release%20-%20Section%20III%20Minimum%20Deterrence%20US%20Nuclear%20Weapons%20and%20Priority%20Threats.pdf.

Onur Bakir, Niyazi, and Detlof von Winterfeldt. "Is Better Nuclear Weapon Detection Capability Justified?" *Journal of Homeland Security and Emergency Management* 8, no. 1 (2011): art. 16. https://doi.org/10.2202/1547-7355.1731.

A Review of U.S. International Efforts to Secure Radiological Materials, Hearing before the Oversight of Government Management, the Federal Workforce, and the District of Columbia Subcommittee of the Committee on Homeland Security and Governmental Affairs, United States Senate, 110th Cong., 1st sess. (March 13, 2007). https://www.congress.gov/event/110th-congress/senate-event/LC9695/text.

Schaeffer, Richard L., William Mendenhall III, R. Lyman Ott, and Kenneth G. Gerow. *Elementary Survey Sampling*. 7th ed. Boston: Brooks/Cole, 2012.

Schenkel, Roland. "Improving Verification: Trends and Perspectives for Research." In *Verifying Treaty Compliance*, edited by Rudolf Avenhaus, Nicholas Kyriakopoulos, Michel Richard, and Gotthard Stein, 589–603. Berlin, Heidelberg: Springer, 2006. https://doi.org/10.1007/3-540-33854-3_29.

Scouras, James, ed. *On Assessing the Risk of Nuclear War*. Laurel, MD: Johns Hopkins University Applied Physics Laboratory, 2021. https://www.jhuapl.edu/sites/default/files/2022-12/OnAssessingRiskNuclearWar.pdf.

"Sen. Lugar Releases New Report on WMD Threats, Responses." *US Fed News Service, Including US State News*, June 22, 2005. https://www.proquest.com/docview/470638708/citation/44BE5A7A95DA47AAPQ/1?.

Sudnik, John. " 'Dirty Bomb' Attack: Assessing New York City's Level of Preparedness from a First Responder's Perspective." Master's thesis, Naval Postgraduate School, 2006. DTIC (accession no. ADA445265). https://apps.dtic.mil/sti/citations/ADA445265.

Taleb, Nassim Nicholas. *The Black Swan: The Impact of the Highly Improbable*. 2nd ed. New York: Random House, 2010.

Trends in Illicit Movement of Nuclear Materials, Hearing before the Subcommittee on Prevention of Nuclear and Biological Attack of the Committee on Homeland Security, House of Representatives. 109th Cong., 1st sess. (September 22, 2005). https://www.govinfo.gov/content/pkg/CHRG-109hhrg31781/html/CHRG-109hhrg31781.htm.

Wirz, Christoph, and Emmanuel Egger. "Use of Nuclear and Radiological Weapons by Terrorists?" *International Review of the Red Cross* 87, no. 859 (2005): 497–510. https://international-review.icrc.org/sites/default/files/irrc_859_5.pdf.

Zadeh, Lotfi A. "Fuzzy Sets." *Information and Control* 8, no. 3 (1965): 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X.

Acknowledgments

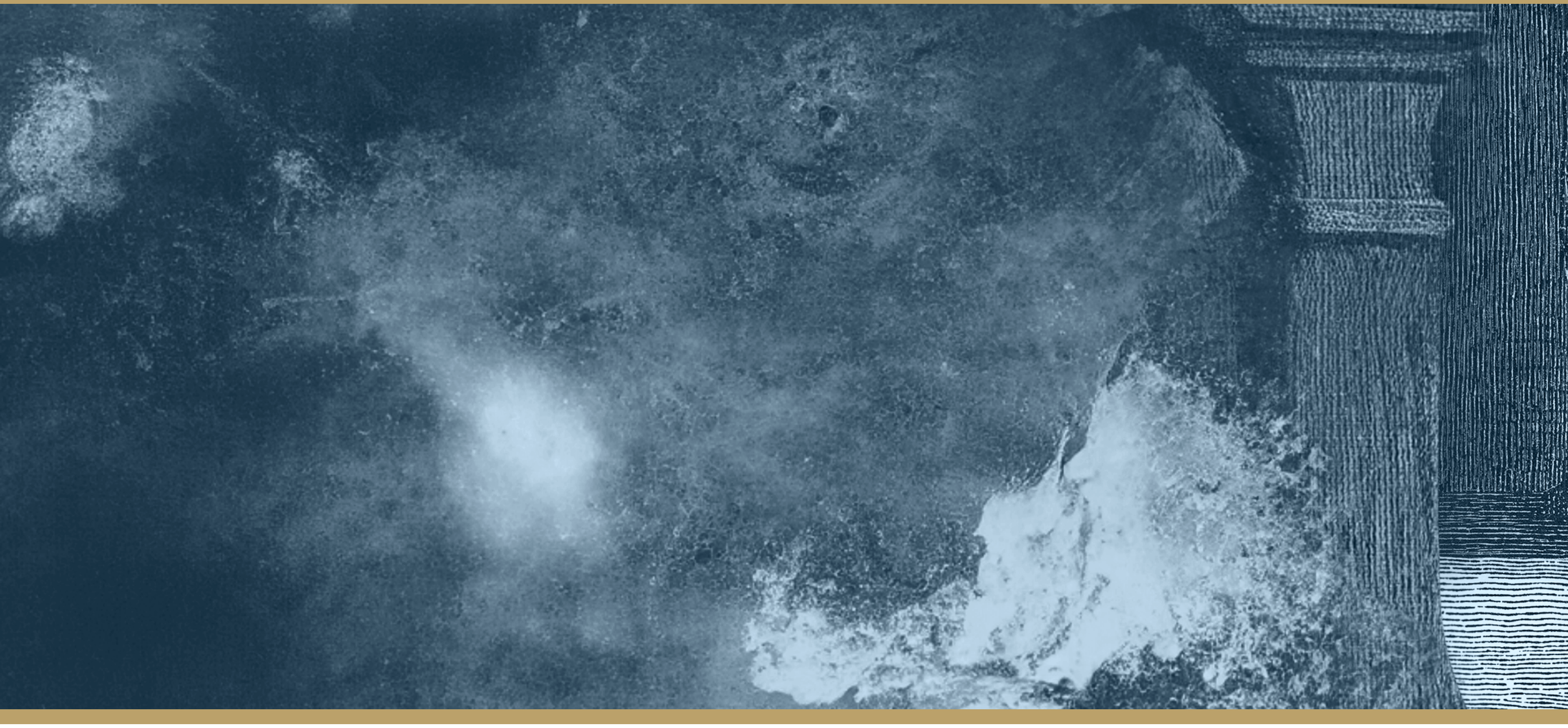
We thank Dan Diller, director of policy at the Lugar Center and former staffer of the late Senator Richard Lugar, for speaking with us and imparting his knowledge and experience about the Lugar survey and its report. We also thank Seth Baum, Dennis Evans, Christian Ruhl, and Scott Sagan for their thoughtful reviews of earlier drafts of this paper.

About the Authors

Jane Booker, currently a consultant, was formerly group leader of the Statistics Group at Los Alamos National Laboratory. She is nationally known for her pioneering work in eliciting and analyzing expert knowledge, uncertainty quantification, statistical reliability, and information integration methods. Among her publications are the books *Eliciting and Analyzing Expert Judgment: A Practical Guide*, coauthored with Mary A. Meyer, and *Fuzzy Logic and Probability Applications: Bridging the Gap*, coauthored with Timothy J. Ross and the late W. Jerry Parkinson. She is a member of the Institute of Mathematical Statistics, a fellow of the American Statistical Association, the 1995 recipient of the H. O. Hartley Award for service to the statistics profession, and recipient of the prestigious R&D 100 Award for the reliability methodology PREDICT. Dr. Booker holds a PhD in statistics from Texas A&M University.

Lori Baxter is an analyst in the Concepts and Assessments Group in the National Security Analysis Department at the Johns Hopkins University Applied Physics Laboratory (APL). While working at APL, she has been involved in projects on strategic deterrence, army modernization, anti-submarine warfare, and nuclear policy. Much of her research involves using qualitative methods to analyze public policy and military operations. She graduated from Johns Hopkins in 2021 with a BS in applied mathematics and statistics and international studies, and she is currently pursuing an MA in global security studies, also from Johns Hopkins.

James Scouras is a senior scholar at the Johns Hopkins University Applied Physics Laboratory (APL) where he leads a research program on nuclear war and other global catastrophic risks. He was formerly chief scientist of the Defense Threat Reduction Agency's Advanced Systems and Concepts Office. Prior to that, he was program director for risk analysis at the Homeland Security Institute, held research positions at the Institute for Defense Analyses and the RAND Corporation, and lectured on nuclear policy in the University of Maryland's General Honors Program. Among his publications are the book *A New Nuclear Century: Strategic Stability and Arms Control* (Praeger, 2002), coauthored with Stephen Cimbala, and his edited volume *On the Risk of Nuclear War* (APL, 2021). Dr. Scouras earned his PhD in physics from the University of Maryland.



APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY