

COMMUNICATING RISK IN INTELLIGENCE FORECASTS:
THE CONSUMER'S PERSPECTIVE

by

NATHAN F. DIECKMANN

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2007

“Communicating Risk in Intelligence Forecasts: The Consumer’s Perspective,” a dissertation prepared by Nathan F. Dieckmann in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology. This dissertation has been approved and accepted by:

Paul Slovic, Chair of the Examining Committee

11/26/07

Date

Committee in Charge: Paul Slovic, Chair
Robert Mauro
Bertram Malle
John Orbell

Accepted by:

Dean of the Graduate School

© 2007 Nathan F. Dieckmann

An Abstract of the Dissertation of
Nathan F. Dieckmann for the degree of Doctor of Philosophy
in the Department of Psychology to be taken December 2007
Title: COMMUNICATING RISK IN INTELLIGENCE FORECASTS: THE
CONSUMER'S PERSPECTIVE

Approved: _____
Paul Slovic

The main goal of many political and intelligence forecasts is to effectively communicate risk information to decision makers (i.e. consumers). Standard reporting most often consists of a narrative discussion of relevant evidence concerning a threat, and rarely involves numerical estimates of uncertainty (e.g. a 5% chance). It is argued that numerical estimates of uncertainty will lead to more accurate representations of risk and improved decision making on the part of intelligence consumers. Little work has focused on how well consumers understand and use forecasts that include numerical estimates of uncertainty. Participants were presented with simulated intelligence forecasts describing potential terrorist attacks. These forecasts consisted of a narrative summary of the evidence related to the attack and numerical estimates of likelihood and potential harm. The primary goals were to explore how the structure of the narrative summary, the format of likelihood information, and the numerical ability (numeracy) of consumers affected perceptions of

intelligence forecasts. Consumers perceived forecasts with numerical estimates of likelihood and potential harm as more useful than forecasts with only a narrative evidence summary. However, consumer's risk and likelihood perceptions were more greatly affected by the narrative evidence summary than the stated likelihood information. These results show that even "precise" numerical estimates of likelihood are not necessarily evaluable by consumers and that perceptions of likelihood are affected by supporting narrative information. Numeracy also moderated the effects of stated likelihood and the narrative evidence summary. Consumers higher in numeracy were more likely to use the stated likelihood information and consumers lower in numeracy were more likely to use the narrative evidence to inform their judgments. The moderating effect of likelihood format and consumer's perceptions of forecasts in hindsight are also explored.

Explicit estimates of uncertainty are not necessarily useful to all intelligence consumers, particularly when presented with supporting narrative evidence. How consumers respond to intelligence forecasts depends on the structure of any supporting narrative information, the format of the explicit uncertainty information, and the numerical ability of the individual consumer. Forecasters should be sensitive to these three issues when presenting forecasts to consumers.

CURRICULUM VITAE

NAME OF AUTHOR: Nathan F. Dieckmann

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon
San Francisco State University

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2007, University of Oregon
Master of Science, Psychology, 2004, University of Oregon
Bachelor of Arts, Psychology, 2001, San Francisco State University

AREAS OF SPECIAL INTEREST:

Judgment & Decision Making
Risk Analysis & Communication
Statistical Analysis

PROFESSIONAL EXPERIENCE:

Research Scientist, Decision Research, Eugene, OR, 2007-present.

Founding Partner, Integrative Analytic Consulting, 2005-present.

Research Associate, Decision Research, Eugene, OR, 2004-2007.

Graduate Teaching Fellow, Department of Psychology, University of Oregon,
2002-2007.

Research Coordinator, Stanford Bipolar Disorders Clinic, 2000-2002.

Research Assistant, Stanford Psychiatry Neuroimaging Laboratory, 2001-2002.

GRANTS, AWARDS AND HONORS:

Finalist, APA Summer Fellowship in DoD counterintelligence, 2006.

Travel award, Annual meeting of the American Psychiatric Association, 2002.

Travel award, Annual meeting of the American Psychiatric Association, 2001.

PUBLICATIONS:

Peters, E., Dieckmann, N.F., Dixon, A., Slovic, P., Mertz, C.K., & Hibbard, J. H. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review*, 64(2), 169-190.

Peters, E., Hibbard, J. H., Slovic, P., & Dieckmann, N. F. (2007). Numeracy skill and the communication, comprehension, and use of risk and benefit information. *Health affairs*, 26(3), 741-748.

Hall, M., Sedlacek, A., Berenbach, A., & Dieckmann, N. F. (in press). Military sexual trauma services for women veterans in the veterans health administration: The patient-care practice environment and perceived organizational support. *Psychological Services*.

Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2007). An empirical assessment of meta-analytic practice. Submitted to *Review of General Psychology*.

Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C.K., Slovic, P., & Hibbard, J. H. (2007). Bringing meaning to numbers: The effects of affect in choice. Submitted to *The Journal of Experimental Psychology*.

Dieckmann, N. F. (2007). Numeracy: A review of the literature. Report submitted to National Cancer Institute.

Mauro, R., Barshi, I., Dieckmann, N. F., & Shepler, C. (2006). TRIAD: Tool for Risk Identification, Assessment, & Display. Report submitted to the National Engineering Safety Council (NESC; NASA).

Ketter, T. A., Wang, P. W., Dieckmann, N. F., Lembke, A., Becker, O. & Camilleri, C. (2003). Brain Anatomic Circuits and the Pathophysiology of Affective Disorders. In J. C. Soares (Ed.), *Brain Imaging in Affective Disorders* (pp.79-118). New York, NY: Marcel Dekker.

ACKNOWLEDGMENTS

The completion of my graduate studies would not have been possible without the supportive mentors and fellow graduate students at the University of Oregon. I would like to thank Paul Slovic, Robert Mauro, Ellen Peters and Bertram Malle for all of their support throughout my 5-year graduate career. Their unique strengths and perspectives have profoundly shaped my development as a researcher and teacher. I am particularly appreciative of their understanding and support of my need for a healthy balance between work, family, and leisure. I would like to thank John Orbell for agreeing to serve on my dissertation committee and for his helpful suggestions concerning my dissertation work. I am also grateful to the entire team at Decision Research for offering me an opportunity to work and play within such a great organization.

My journey through graduate school would certainly have been less fun and less productive without the friendship of Jonathan Cook, Stephan Dickert, and Jess Holbrook. Thanks for everything guys.

I also want to thank my family for always supporting my various life journeys and always trusting and helping me along the way. Finally, I would like to thank my wonderfully supportive wife Rahwa. There is no doubt that without her none of this would have been possible. Thanks so much for being such a loving and supportive partner.

TABLE OF CONTENTS

| Chapter | Page |
|---|------|
| I. INTRODUCTION AND PROBLEM STATEMENT | 1 |
| Introduction to Policy and Intelligence Analysis | 1 |
| Reporting the Results of Intelligence Forecasts | 2 |
| The Benefits of Including Explicit Estimates of Uncertainty | 4 |
| Why Might Analysts be Reluctant to Include Explicit Estimates of Uncertainty? | 5 |
| Summary | 8 |
| II. LITERATURE REVIEW | 10 |
| Uncertainty, Probability, and Sensitivity Analysis | 10 |
| The Interpretation of Probability | 11 |
| The Form of a Probability Statement | 13 |
| Sensitivity Analysis and Presenting Ranges of Plausible Values | 14 |
| Intuitive Probability Judgments | 15 |
| Strategies for Intuitive Probability Judgments | 15 |
| Intuitive Judgments from both Numerical Probabilities and Scenario- based Information | 18 |
| Intuitive Judgments of Risk | 20 |
| Consumer Perceptions of Probabilistic Forecasts and Risk Communications | 21 |
| Effects of the Format of Uncertainty Information on Perceptions of Risk and Perceptions of the Quality of the Forecast or Forecaster | 23 |
| Verbal versus Numerical Expressions of Uncertainty | 23 |
| Standard Probability Formats versus Frequency Formats | 24 |
| Confidence Intervals and Reporting a Range of Plausible Probability Values in a Forecast | 25 |
| Internal versus External Framing of Probabilistic Forecasts | 27 |
| Numeracy | 28 |
| The Conceptualization and Measurement of Numeracy | 28 |

| Chapter | Page |
|--|------|
| Previous Findings Relating to Individual Differences in Numeracy | 30 |
| Numeracy and Affective Processing | 31 |
| Hindsight Bias | 33 |
| | |
| III. A MODEL OF CONSUMER PERCEPTIONS OF SINGLE-EVENT INTELLIGENCE FORECASTS AND PRIMARY RESEARCH QUESTIONS | 36 |
| A Model of Consumer Perceptions of Single-Event Intelligence Forecasts | 36 |
| Research Questions | 38 |
| | |
| IV. THE SIMULATED INTELLIGENCE FORECASTS, THE SUBJECT POPULATION, AND PRELIMINARY STUDY 1 | 41 |
| Development of the Simulated Intelligence Forecasts | 41 |
| Experimental Participants | 42 |
| Preliminary Study 1: Pretesting the Simulated Intelligence Forecasts | 42 |
| | |
| V. PRIMARY EXPERIMENTAL STUDIES | 47 |
| Overview of Primary Studies..... | 47 |
| Study 2 – Initial Explorations of the Effects of Explicit Likelihood and Scenario Information on Perceptions of Risk..... | 47 |
| Purpose | 47 |
| Method..... | 48 |
| Results | 50 |
| Summary and Discussion | 65 |
| Study 3 – Further Investigations of Including Numerical Estimates of Likelihood and Harm in Forecasts | 68 |
| Purpose..... | 68 |
| Method..... | 70 |
| Results | 73 |
| Summary and Discussion | 92 |
| Study 4 – Exploring Consumer Perceptions of Intelligence Forecasts in Hindsight..... | 95 |

| Chapter | Page |
|---|------------|
| Purpose..... | 95 |
| Method..... | 96 |
| Results | 98 |
| Summary and Discussion | 109 |
| VI. CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH DIRECTIONS | 113 |
| Conclusions and Implications | 113 |
| Perceptions of Intelligence Forecasts with Numerical Likelihood and Narrative Information | 113 |
| The Formatting of Numerical Likelihood in Intelligence Forecasts | 117 |
| Individual Differences in the Numerical Ability of Consumers | 119 |
| Limitations | 121 |
| Future Research Directions | 122 |
| APPENDICES | 125 |
| A. SCENARIOS TESTED IN PRELIMINARY STUDY 1 | 125 |
| B. MATERIALS FOR STUDY 2 | 132 |
| C. MATERIALS FOR STUDY 3 | 140 |
| D. MATERIALS FOR STUDY 4..... | 152 |
| E. DETAILS OF STATISTICAL ANALYSES | 161 |
| REFERENCES..... | 171 |

LIST OF FIGURES

| Figure | Page |
|---|------|
| 1. A simple model of consumer risk perceptions from simulated intelligence forecasts | 37 |
| 2. The effect of stated likelihood and likelihood format on perceived risk | 55 |
| 3. The effect of format condition and numeracy level on perceptions of risk | 56 |
| 4. The effect of stated likelihood and likelihood format on perceived risk with summary only, for consumers low in numeracy | 57 |
| 5. The effect of stated likelihood and likelihood format on perceived risk with summary only, for consumers high in numeracy | 58 |
| 6. The effect of stated likelihood and likelihood format on perceived risk with summary plus evidence, for consumers low in numeracy | 58 |
| 7. The effect of stated likelihood and likelihood format on perceived risk with summary plus evidence, for consumers high in numeracy | 59 |
| 8. The effect of uncertainty format and consumer numeracy on perceptions of usefulness, knowledge and trust | 62 |
| 9. The effect of uncertainty format and numeracy on perceptions of usefulness, knowledge and trust at low stated likelihood (i.e. 5%, 5/100) ... | 63 |
| 10. The effect of uncertainty format and numeracy on perceptions of usefulness, knowledge and trust at high stated likelihood (i.e. 20%, 20/100)..... | 64 |
| 11. The effect of uncertainty format and stated likelihood on perceptions of usefulness, knowledge and trust | 65 |
| 12. Boxplots showing the distribution of likelihood ratings at each level of likelihood and likelihood format | 78 |
| 13. The effect of stated likelihood on perceived likelihood for the point estimate and range conditions..... | 84 |
| 14. The relationship between perceived credibility/coherence and perceived likelihood for consumers in the point estimate and range conditions | 87 |
| 15. The effect of stated likelihood on perceived likelihood for consumers with different levels of numeracy | 88 |
| 16. The relationship between perceived credibility/coherence and perceived likelihood for consumers with different levels of numeracy | 89 |

| Figure | Page |
|--|------|
| 17. The distributions of hindsight likelihood ratings by stated likelihood and likelihood format. | 108 |

LIST OF TABLES

| Table | Page |
|--|------|
| 1. Percentage of participants that felt each probability was a reasonable estimate of the probability that the event would occur | 45 |
| 2. Descriptive statistics for subject-generated probability ranges | 46 |
| 3. Sample Characteristics | 50 |
| 4. Sample Characteristics | 51 |
| 5. Pearson correlations (w/ 95% CI) between dependent variables | 51 |
| 6. The effect of explicit likelihood estimates and uncertainty format on risk perceptions <u>without</u> a narrative description of evidence | 53 |
| 7. The effect of explicit likelihood estimates and uncertainty format on risk perceptions <u>with</u> a narrative description of evidence | 54 |
| 8. Sample Characteristics | 74 |
| 9. Sample Characteristics | 74 |
| 10. Average Pearson correlations w/ 95% CI's between dependent variables related to risk perception | 75 |
| 11. The effect of explicit numerical estimates of likelihood/harm and uncertainty format on consumer perceptions of likelihood | 77 |
| 12. Multilevel model results for perceived likelihood..... | 86 |
| 13. The effect of stated likelihood and likelihood format on perceived usefulness..... | 91 |
| 14. The effect of stated likelihood and likelihood format on perceived source credibility | 92 |
| 15. Sample Characteristics..... | 98 |
| 16. Sample Characteristics..... | 98 |
| 17. Average Pearson correlations w/ 95% CI's between blame, usefulness, and source credibility..... | 99 |
| 18. Average Pearson correlations w/ 95% CI's between dependent variables related to risk perception | 100 |
| 19. The effect of stated likelihood and likelihood format on perceptions of blame..... | 102 |
| 20. The effect of stated likelihood and likelihood format on perceived usefulness..... | 104 |

| Table | Page |
|---|------|
| 21. The effect of stated likelihood and likelihood format on perceived source credibility | 105 |
| 22. The effect of stated likelihood and likelihood format on perceptions of likelihood | 107 |

CHAPTER I INTRODUCTION AND PROBLEM STATEMENT

Introduction to Policy and Intelligence Analysis

Human societies are more interdependent with respect to economics, culture, and human ecology than at any other time in history. As a result of this increased dependency, political and business leaders are often faced with monumental decisions about policy that have the potential to affect vast numbers of people around the world. Thus, it is very important that these decisions are based on the very best information and analysis.

There are numerous public and private agencies that conduct analysis and research with the goal of aiding political decision makers. This decision support is called policy analysis and/or policy focused-research. Morgan & Henrion (1990) define policy analysis as an “analytical activity undertaken in direct support of specific public or private sector decision makers who are faced with a decision that must be made or a problem that must be resolved” (pg. 16). In addition, “the objective of policy analysis ‘is to evaluate, order and structure incomplete knowledge so as to allow decisions to be made with as complete an understanding as possible of the current state of knowledge, its limitations and implications’ (Morgan, 1978)”.

US intelligence agencies are an example of a public entity that provides policy analysis and policy-related research to senior US decision makers. The analysis and forecasting activities of US intelligence agencies take several unique forms, which can be categorized into three basic types of intelligence – strategic, tactical, and indications and warnings intelligence (Clark, 2004; a similar categorization is discussed by Cooper, 2005). Strategic intelligence involves in-depth research focused on the capabilities and plans of a target. These are long-range intelligence products that tend to be broad and complex in terms of both information sources and the time window covered in the report.

The National Intelligence Estimates (NIE's) generated by the Central Intelligence Agency (CIA) are strategic intelligence products. In contrast to the complexity and breadth of the NIE's, tactical intelligence involves the collection and transmission of current (real-time) information to support issues that require immediate action or are currently being executed. Finally, arguably the highest priority activity for an intelligence agency is "providing indications and warning on threats to national security" (Clark, 2004, pg. 159). Indications and warnings intelligence involves "detecting and reporting time-sensitive information on foreign developments that threaten the country's military, political, or economic interests" (Clark, 2004, pg. 159). Generating forecasts and providing reports warning of potential terrorist attacks is one example of this type of intelligence product.

One of the primary goals of indications and warnings intelligence is communicating risk information to government decision makers or other consumers of the risk analyses¹. As Fisk (1995) comments, "Problems of 'indications analysis' or 'intelligence warning' are essentially questions of how to assign probabilities to hypotheses of interest" (pg. 264). For example, one hypothesis of interest could be the proposition that a known terrorist group will carry out a specific terrorist act within a given timeframe. Decision makers responsible for national security would greatly benefit by being warned of such a plot, and ideally they would also like to know the chances that this attack will occur and the potential harm that would result if the attack were to succeed. In this sense, the indications and warnings intelligence process is really a form of estimating and communicating risks. The present work is focused on understanding the factors that influence consumer perceptions of indications and warnings intelligence products.

Reporting the Results of Intelligence Forecasts

US intelligence analysts, and policy analysts more generally, have traditionally relied on qualitative methods for the bulk of their analysis and forecasting. Two main

¹ In this context, the "consumers" are those individuals that use intelligence forecasts to make decisions about policy and action (e.g. military leaders, the president, and others).

techniques include model based approaches (see Clark, 2004) and scenario-based forecasting (see Clark, 2004; Schwartz, 1996). One common feature of these qualitative approaches is that probability and/or margins of uncertainty are not explicitly represented when developing and reporting forecasts. Although there are several modern examples of large-scale policy analyses in which numerical uncertainties are estimated and reported (Morgan & Henrion, 1990), explicit uncertainty analysis has not reached the state of standard practice in many domains of policy and risk analysis.

Several authors have noted that the insufficient description of probability and analyst uncertainty has contributed to intelligence failures and other difficulties in communicating forecasts. For example, Armstrong, Leonhart, McCaffery & Rothenberg (1995) discuss several intelligence failures that were caused, at least in part, by a “reluctance to quantify their [the analysts’] theories of probability or their margins of uncertainty” (pg. 240). The historical forecasts they examined included the first Chinese nuclear test, the OPEC price decrease of December 1973, and the Ethiopian revolution of 1974. In addition, Michael Schrage in an editorial for the Washington Post (February, 20th, 2005) discussed the importance of analysts including estimates of uncertainty in intelligence reports. He describes the lack of quantitative uncertainty estimates as an institutional bias and points out that many other professionals, including insurance analysts, bankers and public health practitioners, routinely use quantitative risk analyses. Why should intelligence forecasts concerning national security, often reported directly to the President, have less analytic complexity than the forecasts generated by the professionals mentioned above? The closest that most intelligence analysts come to quantifying probability or margins of uncertainty are vague verbal probability estimates (i.e. this attack “could” occur; the attack is “highly unlikely” at this point, etc; Zlotnick, 1995). However, because verbal probability statements are poorly defined and may mean different things to different people, they are not ideal for the accurate communication of risk (Armstrong et al., 1995; see Chapter II).

When it comes to communicating the results of an intelligence forecast, most finished intelligence products are presented in scenario-based or narrative form that describe the

possible future states of the world or target (Clark, 2004). Analysts appear to “prefer to transmit knowledge through writing, because only writing can capture the full complexity of what they have to say” (Gardiner, 1995, pg.354). Consequently, the form and style of these narrative reports is an important part of communicating risk and analytic conclusions between analyst and consumer. In fact, several authors have discussed methods for writing convincing scenarios to increase the chances that consumers will accept a forecast (Clark, 2005; Gregory & Duran, 2001). However, because of institutional norms and consumers’ preferences for information in narrative form, intelligence reports are not likely to become purely quantitative in nature. The most natural way to include quantitative estimates of uncertainty in current intelligence reporting is along side supporting narrative information concerning the evidence and reasoning supporting the conclusions.

The Benefits of Including Explicit Estimates of Uncertainty

Morgan and Henrion (1990) discuss several general reasons why explicitly addressing uncertainty is important in policy and intelligence analysis. Their first argument is one by analogy, arguing that if natural scientists are expected to be explicit about uncertainty in measured quantities, why shouldn’t policy-focused research be held to the same standard, particularly because the uncertainty is much greater in the policy domain than in the natural sciences? They also point out three more specific arguments in favor of uncertainty analysis and explicit reporting:

1. A central purpose of policy research and policy analysis is to help identify the important factors and the sources of disagreement in a problem, and to help anticipate the unexpected. An explicit treatment of uncertainty forces us to think more carefully about such matters, helps us to identify which factors are most and least important, and helps us plan for contingencies or hedge our bets.
2. Increasingly we must rely on experts when we make decisions. It is often hard to be sure we understand exactly what they are telling us. It is harder still to know what to do when different experts appear to be telling

us different things. If we insist they tell us about the uncertainty of their judgments, we will be clearer about how much they think they know and whether they really disagree.

3. Rarely is any problem solved once and for all. Problems have a way of resurfacing. The details may change but the basic problems keep coming back again and again. Sometimes we would like to be able to use, or adapt, policy analyses that have been done in the past to help with the problems of the moment. This is much easier to do when the uncertainties of the past work have been carefully described, because then we can have greater confidence that we are using the earlier work in an appropriate way. (pg. 3)

Related to the third point above, Fisk (1995) and Schrage (2005) note that consistently including quantitative uncertainty estimates in intelligence reports could act as an audit trail for analytic judgment, which could be revisited and reviewed by consumers and the analytic community. Schrage (2005) also points to several other benefits of this greater analytic accountability. For one thing, it would put pressure on analysts to think extra hard about their analysis and conclusions. It would also give consumers much more information on which to judge the analytic conclusions, and ideally, the explicit uncertainty estimates would allow a more accurate transferal for risk information from analyst to consumer (e.g. the likelihood that that the analyst assigns to the potential threat is accurately communicated to the consumer). Consumers could quickly assess the level of confidence that an analyst has in his or her evidence and conclusions, and the explicit uncertainties would give consumers an idea of where more work needs to be focused to reduce the uncertainty: “Then their ability to push, prod and poke the intelligence community would be firmly grounded in their own perception of the strength and weakness of the work coming out of it” (Schrage, 2005).

Why Might Analysts be Reluctant to Include Explicit Estimates of Uncertainty?

The analytic community continues to primarily focus on qualitative techniques for forecasting in which they do not consistently provide numerical estimates of uncertainty. There are several reasons why this may be the case. First, many intelligence problems

are so complex and multifaceted, and involve such a great degree of uncertainty, that it may seem impossible to estimate the uncertainty in the system. Morgan and Henrion (1990) note that it may be because of the “vast uncertainties” inherent in many policy analyses that it is “still not standard practice to treat uncertainties in an explicit probabilistic fashion” (pg.20). Considering this enormous complexity and uncertainty, it is not surprising that many analysts see a qualitative approach to analysis as the only alternative. The work of the intelligence analyst has even been compared to that of the historian, both of which labor to fit disparate pieces of evidence together into a coherent causal story (Heuer, 1999). For many intelligence problems, this focus on forming a coherent story out of a set of evidence may lead analysts into a scenario/narrative presentation of the results and away from thinking probabilistically about their conclusions and forecasts.

It is clear that there are many situations in which uncertainty must be estimated through expert judgment alone, and it is understandable that this may seem like a daunting task. However, most analysts would agree that they cannot be sure about the level of uncertainty present in a system (for example, the precise probability that an event will occur), but they are not completely ignorant either. Analysts are likely to have some idea or intuition about uncertainty, and there are several structured techniques that can be used to help elicit probability estimates from experts (see Armstrong, 2001).

The second reason that analysts may be reluctant to use numerical estimates of uncertainty in forecasts is that they feel that there are no structured techniques available for applying risk analysis or estimating uncertainties in the intelligence domain. In recent years, however, several different schemes and approaches for probabilistic and uncertainty analysis that would be applicable to intelligence problems have been developed. For example, several authors have discussed the potential application of quantitative risk analysis procedures to problems of terrorism prediction and forecasting (for example see Garrick, 2002; Pate-Cornell, 2002; Haines and Longstaff, 2002; Horowitz & Haines, 2003).

A third reason is that analysts may feel that even if they did explicitly report probability and margins of error in intelligence reports, consumers would not be interested in seeing them, nor would they be able to understand or use the information. Michael Schrage in his Washington Post editorial (February, 20th, 2005) relates a conversation that he had with a senior CIA officer concerning consumers and quantitative analyses: “Intelligence analysts ‘would rather use words than numbers to describe how confident we are in our analysis,’ a senior CIA officer who's served for more than 20 years told me. Moreover, ‘most consumers of intelligence aren't particularly sophisticated when it comes to probabilistic analysis. They like words and pictures, too. My experience is that [they] prefer briefings that don't center on numerical calculation. That's not to say we can't do it, but there's really not that much demand for it.’”

It is an empirical question as to how well consumers, particularly those uncomfortable with numbers, would be able to use, and feel comfortable using, intelligence forecasts that include quantitative estimates of uncertainty. There is a relatively rich psychological literature on how people perceive likelihood and risk, and the experimental work in this dissertation will focus on exploring lay consumers' perceptions of forecasts in the intelligence domain.

A last potential concern is that providing explicit estimates of uncertainty would leave an audit trail of analytic forecasts. Analysts may be reluctant to leave themselves open to potential criticism if events to which they assign small probabilities occur, or events to which they assign high probabilities do not occur. It may be more comforting to keep analytic judgments and forecasts vague, which allows only “ambiguous accountability” (Schrage, 2005). This may be partly a fear about hindsight bias on the part of future auditors of an analyst's forecasts, as well as a fear about finding out how poorly calibrated their forecasts really are (see Tetlock, 2005; Heuer, 1999).

It is unknown how consumers will feel about analytic judgments in hindsight. For example, will consumer perceptions be greatly affected by the presence of explicit uncertainties in intelligence forecasts? Another focus of the empirical work in this dissertation is on how consumers will view the results of quantitative forecasts in

hindsight.

The first two issues concerning analytic methods are outside of the scope of the present work. The empirical work in this dissertation will focus on perceptions of risk forecasts from the perspective of the consumer.

Summary

The focus of this dissertation is on indications and warnings intelligence forecasts. The purpose of these forecasts is to communicate risk information in a format that is effective and subsequently useful for decision making. However, standard reporting methods in policy and intelligence analysis rarely involve explicit, numerical estimates of uncertainty. Even though there are many potential benefits of including numerical uncertainty estimates in policy and intelligence forecasts, the analytic community has been reluctant to express uncertainty in quantitative form. Standard reporting methods for intelligence forecasts most often involve a scenario-based or narrative discussion of the evidence and possible future states of the world, and any numerical estimates of uncertainty would likely accompany this narrative presentation.

Several writers have argued that the explicit treatment of uncertainty will lead to improved analysis and risk communication (e.g. Morgan and Henrion, 1990; Schrage, 2005). Quantitative estimates of the likelihood and potential harm of particular target events (ideally with an accompanying sensitivity analysis) may lead consumers of intelligence forecasts to more accurately perceive the attendant risks and to make better decisions. One potential benefit of including quantitative estimates in intelligence forecasts is greater consistency in interpretations.

However, two things must happen for this quantitative approach to improve consumer decision making. First, the analysts must use solid analytic methods and reach sound conclusions. As discussed briefly above, several risk and policy analysts have developed techniques for conducting quantitative risk analyses in the intelligence domain. Second, the consumers of these reports must be able to understand and be comfortable using the results of these quantitative analyses. If consumers misinterpret the results, or otherwise

misuse or ignore them, then the hard work done by the analysts is lost. It is clear that the communication between analysts and consumers is critical component of the process. Although many authors have discussed particular analytic techniques that could be fruitfully applied in the intelligence domain, how these analyses should be reported for the benefit of consumers has received less attention.

In an intelligence forecast that includes both scenario-based and numerical uncertainty information, there are several sources of information that consumers can use to make judgments of risk and quality. The focus of this dissertation is on risk communication, specifically on how consumers understand and evaluate quantitative intelligence forecasts concerning the risk of terrorist attacks. The primary goals are to explore how the structure and format of an intelligence forecast, as well as the individual characteristics of the consumer (e.g. a consumer's ability to understand probability information), affect consumer perceptions of risk and perceptions of the usefulness and quality of intelligence forecasts. Another aim of this work is to model how consumers use these various sources of information to inform their judgments.

Selected research literature related to the conceptualization of uncertainty, intuitive perceptions of likelihood and risk, individual differences in numerical ability, and the effect of hindsight knowledge is reviewed in the next chapter. In Chapter III, a model of consumer risk perception is developed along with specific research questions for the empirical work that follows. The following chapters consist of the experimental results and conclusions, as well as the implications of this work for the communication of risk in indications and warnings intelligence forecasts.

CHAPTER II LITERATURE REVIEW

Uncertainty, Probability and Sensitivity Analysis

The explicit representation of uncertainty is important for policy and intelligence forecasting, but how can we conceptualize and define uncertainty, probability, and sensitivity analysis? According to Rowe (1994), “Uncertainty is essentially the absence of information, information that may or may not be obtainable.” (pg. 743). In general, when analysts are asked to report their uncertainty in a forecast, they are being asked to detail or quantify the effect that imperfect information has had on the results of the analysis. This type of uncertainty has also been called epistemic uncertainty, which is conceptually different from aleatory uncertainty (Pate-Cornell, 1996). Aleatory uncertainties “stem from variability in known (or observable) populations and, therefore, represent randomness in samples”, and epistemic uncertainties stem “from a basic lack of knowledge about fundamental phenomena” (pg. 97). Most problems in risk, policy, and intelligence analysis will involve both types of uncertainty, although epistemic uncertainty will tend to dominate.

Rowe (1994) describes four different classes of uncertainty important in risk analyses: 1) Metrical, uncertainty and variability in measured quantities; 2) Structural, uncertainty due to complexity in modeling the phenomenon under study; 3) Temporal, uncertainty about future and past states of the world, and 4) Translational, uncertainty in transmitting information through the explanation of uncertain results (see also Politi, Han & Col, 2006; Peters, 2006). Metrical uncertainty is extremely important in intelligence/policy analysis, as it is directly related to the quality and credibility of the evidence on which an analysis is based. Evidence credibility is not only important in the analysis stage, but might also be helpful to include in a final report for consumers (Heuer, 1999; Schrage, 2005). For instance, information about evidence credibility could help

consumers identify gaps in knowledge that could lead to future information collection efforts. Structural uncertainty is also extremely important because it represents uncertainty in how a model of a phenomenon is constructed. This is a particularly acute problem in policy/intelligence analysis because a large part of the analytic process involves attempts to deduce a model of the situation/target under study. Morgan & Henrion (1990) argue that uncertainty about model form is generally harder to think about than the individual quantities in the model, and that most analysts agree that uncertainty about model form is generally more important and will have large effects on the eventual results and conclusions. Although this may be difficult in practice, ideally analysts would also present a rating of the structural uncertainty in their model. The most familiar kind of uncertainty discussed above is temporal, specifically uncertainty about future states of the world. This type of uncertainty is most often modeled by probability, and this will be a main focus in the empirical studies discussed below.

The Interpretation of Probability

When a person is asked to interpret the meaning of a probability statement (or assess the likelihood of an event), how do they conceptualize “probability”? What does probability mean exactly? This has proven to be a very difficult question, and the collective answer seems to be that it depends.

There are two basic schools of thought about the interpretation of probability: the classical or frequentist school and the subjectivist or Bayesian school. “The classical or frequentist view of probability defines the probability of an event’s occurring in a particular trial as the frequency with which it occurs in a long sequence of similar trials” (Morgan & Henrion, 1990, pg. 48). Thus, probability is only definable if one can locate or generate (at least in principle) a distribution of identical trials of the phenomenon in question (Pate-Cornell, 1996). This means that many of the phenomena to which we assign probabilities, like the probability of a single event occurring, are meaningless from the frequentist point of view. The Bayesian or subjectivist view of probability “is the degree of belief that a person has that it [an event] will occur, given all of the relevant

information currently known to that person” (Morgan & Henrion, 1990, pg. 49). Since a Bayesian probability is by definition subjective and personal, different people may legitimately have different probabilities for the same event, which will depend on their state of knowledge. In practice, when one takes a Bayesian stance toward probabilities, one can incorporate both frequentistic (or aleatory) information about a process or event as well as any other relevant knowledge. In the limited case where only frequency information is available, the subjective Bayesian probability will equal the frequentist probability. This distinction between the objective probabilities based on frequencies and subjective probabilities based on personal belief also roughly maps onto the concept of external and internal statements of probability, respectively (Kahneman & Tversky, 1982a; Teigen, 1994). Frequency based probabilities are restrictively thought of as external to the observer, as a property of the system or process in question, while subjective probabilities include personal statements of uncertainty that are a property of the knower, not a property of the outside world. Some authors have pointed out that this distinction between internal and external probabilities may be part of the gulf between analysts and consumers in terms of communicating risk (Walker, 1995). For example, consumers that adopt a more internal or subjectivist view of likelihood may misunderstand or ignore risk information based on relative frequency interpretations of likelihood.

Several authors have discussed more detailed taxonomies of how probability is understood and interpreted. For example, Teigen (1994) discusses six different interpretations of intuitive probability. Chance probabilities (or Type I) are external and are most naturally thought of in terms of relative frequency. Figuring out the probability of being dealt three of a kind in poker (5-card draw) is a good example of a chance probability. This type of probability is naturally expressed in a frequency format (e.g. 2/100 chance). When thinking about the probability of a single, unique event like a terrorist attack, it becomes more difficult to think of the probability in terms of long-run frequencies (What is the relevant distribution for figuring the frequentist probability?). In these cases, people have been found to rely on different interpretations of probability.

Dispositional probabilities (Type II) are external and can be thought of as a measure of how “easily the outcome in question may occur, or how close it is to becoming realized” (Teigen, 1994, pg. 220). For single unique events people often think of probability as being attached to the event and not as a function of long-run frequencies. Related to dispositional probabilities are their internal counterpart, confidence or subjective degree of belief (Type III; Teigen, 1994). The main issue here is the extent to which a judge believes that the outcome will occur or will soon become realized.

Teigen (1994) also discusses uncertainty by ignorance (Type IV), which concerns not the probability of the chosen hypothesis, but one’s certainty about which hypothesis or prediction to choose in the first place. The next interpretation involves the controllability of events (Type V). For example, personal control may give a sense of certainty that is different than when an event is subject to external, uncontrollable forces. The last variant of the probability concept is plausibility (Type VI). This interpretation of probability is related to perceived closeness to reality, or perceived closeness to truth, and is often activated when one reads a narrative concerning an event. The plausibility, and hence the probability, can be affected by the completeness of the description, the coherence of the story, the causal elements that are included in the story, etc. In practice, quantitative probability estimates will almost always be accompanied by narrative summaries. Hence, the factors that affect plausibility judgments may strongly influence consumer’s perceptions of the probability of events.

The Form of a Probability Statement

Because the focus of this dissertation is on consumer perception of probabilistic forecasts, how analysts format probabilistic information for consumers is very important. A probability can be expressed in percentage form (10%), decimal form (.10) or frequency form (1/10). Each of these forms is mathematically equivalent and, ideally, would be interpreted in the same way. However, as will be discussed further below, the format of the probability information has been found to affect perceptions of likelihood and risk.

Sensitivity Analysis and Presenting Ranges of Plausible Values

Due to the large uncertainties present in many policy or intelligence problems, it is often very difficult for an analyst to generate a point estimate for all empirical quantities and be confident about the structure of the model under study. For example, although it may be difficult to produce a single probability value for the occurrence of an event, analysts can often produce a range of plausible probabilities values. This can be done by first producing a best estimate of the probability and then choosing a high estimate and a low estimate that defines the range of plausible values. As analysts become more confident in their estimates, the confidence interval between the high and the low estimates will become smaller. As analysts become less confident in their estimates, the confidence interval will become larger. For instance, if an analyst was trying to estimate the probability of an event occurring, he or she could report the probability as a range (Low: 10% Best: 25% High: 40%).

Additionally, when structured analytical techniques are used to help an analyst generate a probability value for an event, sensitivity analysis can be used to produce the confidence intervals. Sensitivity analysis refers to changing the inputs, assumptions, or data in an analysis to see how these changes affect the output. There are many different structured techniques that have been developed for sensitivity analysis (see Helton, 1993). For instance, an analyst developing a forecast for a particular terrorist plot could use the worst-case assumptions of the world to get the high probability estimate and then use the best-case assumptions of the world to get the low probability. Producing confidence intervals instead of single point estimates of the probability of an event has the advantage of giving information about the amount of confidence that an analyst has in his or her forecast. Point estimates of the probability of an event will often appear precise regardless of the confidence that an analyst has in the estimate.

In summary, there are several ways of conceptualizing the uncertainty in a policy or intelligence analysis, one of which is temporal uncertainty, or the probability of something happening over a given time frame. The focus in this dissertation is on

consumer perceptions of probability statements, as well as second-order uncertainty, which can be represented by confidence intervals around probability statements. Specifically, I will be focusing on perceptions of intelligence reports that include the constituent pieces of a risk analysis, which is a probability and potential harm estimate.

Intuitive Probability Judgments

A rich psychological literature focuses on how people make intuitive probability judgments about uncertain events. These judgments are intuitive in the sense that they are made without statistical information about the frequency of the target event in the population. For instance, if consumers were presented with a narrative summary of a potential terrorist plot, without explicit estimates of probability, they would need to use intuitive processes to assess the probability or risk of the potential attack. Much of the early work in the field of judgment and decision making focused on intuitive judgments of probability, and it was this work that culminated in the heuristics and biases approach to studying human judgment and reasoning (see Kahneman, Slovic & Tversky, 1982).

Strategies for Intuitive Probability Judgments

Researchers have explored several different strategies that judges use to make unaided intuitive probability judgments when presented with simple descriptions, sets of evidence, or scenarios related to a target event. Two cognitive shortcuts that came out of this literature are using representativeness and availability to judge the likelihood of an uncertain event (Tversky & Kahneman, 1974). Judges use representativeness when they assess the probability of event A by how representative, or how similar, it is of class or process B. For example, judges may estimate the probability that an individual is a member of a particular group by how well the description of that individual resembles their notions of the properties of the group, as in the famous Linda problem (Tversky and Kahneman, 1982ca). In the Linda problem, judges are presented with a short narrative description of the personality and interests of a woman named Linda. They are then asked to choose whether they think Linda is more likely to be a bank teller or a feminist

bank teller (among other options). The narrative description included details about Linda that seem consistent with Linda being a feminist (e.g. deeply concerned about issues of discrimination and social justice), and consequently, the majority of judges thought that Linda was more likely to be a feminist bank teller as opposed to just a bank teller. However, it is clear that the conjunction of two events (i.e. bank teller and feminist) cannot be more likely than a single event (i.e. bank teller), and the judges were said to have succumbed to the conjunction fallacy. Thus, the highly representative description of Linda was thought to have overwhelmed the probabilistic reasoning of the judges. Other biases such as base-rate neglect are also thought to be caused by representativeness, in that judges tend to ignore base-rates when given highly representative scenarios. Much of the experimental work on the representativeness heuristic has been attacked on methodological grounds (e.g. Gigerenzer, 1996), although the notion of representativeness provides one powerful explanation for the robust effects of presenting detailed scenarios of forecasted events on perceptions of likelihood.

We find no good reason to believe that the judgments of political analysts, jurors, judges, and physicians are free of the conjunction effect. This effect is likely to be particularly pernicious in the attempts to predict the future by evaluating the perceived likelihood of particular scenarios. As they stare into the crystal ball, politicians, futurologists, and laypersons alike seek an image of the future that best represents their model of the dynamics of the present. This search leads to the construction of detailed scenarios, which are internally consistent and highly representative of our model of the world. Such scenarios often appear more likely than less detailed forecasts, which are in fact more probable ... The reliance on representativeness, we believe, is a primary reason for the unwarranted appeal of detailed scenarios and the illusory sense of insight that such constructions often provide. (Tversky & Kahneman, 1982a, pg. 97-98)

If judges estimate the likelihood of an event by the “ease with which instances or occurrences can be brought to mind” they have been described as using the availability heuristic (Tversky & Kahneman, 1974, pg. 1128). For example, a judge may rate the likelihood of a particular terrorist attack by the ease with which similar events can be

brought to mind. In this sense, availability is really about probing memory for similar instances with which the judge can use to estimate likelihood. However, Kahneman & Tversky (1982b) also discuss the availability heuristic in terms of the ability to construct instances or scenarios, and they call this the simulation heuristic. In other words, judges may construct plausible scenarios that would lead to the target event and use the ease with which this can be done as a guide to estimating the probability of the event. In the case of the intelligence consumer, this scenario construction is often already completed, and consumers will likely use the “goodness” of the provided narrative scenario as a guide to likelihood estimation. Consumers may, however, intuitively construct additional instances and scenarios from the evidence set.

Several researchers have presented additional models that focus on reasoning as a primary process involved in making intuitive probability judgments. Pennington & Hastie (1988) developed an influential model of explanation-based decision making, which focuses on story construction as the primary reasoning process that mediates many judgments and decisions. The decision maker begins by constructing a mental model (i.e. story, scenario, explanation, or causal model) of the situation from the available evidence. When several potential mental models are reasonable, the best model is chosen based on the fit between the evidence and the story model, as well as by the quality of the story. The perceived quality of the story is determined by the completeness of the explanation, the coherence of the story, the ease of story construction, and other factors (Hastie & Dawes, 2001). This type of scenario-based reasoning strategy seems particularly applicable to the intelligence consumer, in that the consumer would likely use a strategy such as this to make likelihood judgments from the evidence scenario presented by the analyst.

Curley & Benson (1994) discuss a model of belief processing that explicitly focuses on the role of reason construction in likelihood estimation. In their view, probability assessment is more of a reasoning process in which we construct different reasons for or against a proposition (e.g. whether a terrorist attack will occur), form a belief, and then we scale the strength this belief to a probability scale. Tversky & Kahneman (1982b)

also discuss the importance of causal-based reasoning in judgments under uncertainty, in which people are thought to use schemas of cause-effect relationships to make sense of a set of evidence, which is then used as a basis for judgment.

In each of models discussed above, the primary reasoning process involves causal-based scenario construction from a set of evidence on which likelihood judgments are based. These types of reasoning processes can also broadly be classified as knowledge-based as opposed to statistical reasoning. Beach & Braun (1994) discuss a contingency model of subjective probability judgment, in which a judge is thought to possess several different strategies for making probability judgments (e.g. causal-based, statistical, etc), and judges choose the appropriate strategy depending on the context of the problem. For example, with problems that involve games of chance a judge will likely choose to reason statistically, but if given a personality description of an individual in the form of a narrative they are likely to use knowledge-based reasoning strategies. For our purposes, this type of model is interesting in light of the judgment task that intelligence consumers face, in which both scenario-based and explicit probability information is presented in a forecast. In these situations, there may be a conflict between the likelihood estimates based on the scenario presented and the explicit probability presented by the analyst. In this case, consumers are explicitly presented with a numerical probability that is purportedly based on the evidence presented and the professional judgment of the analyst, together with evidence that could be used to create other scenarios and likelihood judgments.

Intuitive Judgments from both Numerical Probabilities and Scenario-based Information

Relatively few studies have explored judgments when both explicit probability and scenario-based information is available to the judge. However, a few researchers have found that scenario information accompanying a probability estimate can have a large effect on the interpretation of that estimate. Windschitl and colleagues (1999, 2002, 2003) have reported several experiments in which they demonstrate that although numerical probability estimates are less affected by context than verbal probability

estimates, they are not free from contextual effects: "...that any numeric probability – whether it is a communicated forecast, an internal belief regarding the objective likelihood of an event, or external information on which a belief is based – can be ambiguous from an intuitive perspective even though it is numerically precise" (Flugstad & Windschitl, 2003, pg. 108).

For example, Flugstad & Windschitl (2003) report several experiments in which participants read scenarios about a doctor's diagnosis that was also accompanied by a numerical estimate of the probability that the surgical intervention would fail. Participants were then asked a series of questions about intuitive optimism or pessimism regarding surgery. The main finding was that, given a fixed numerical probability, positive reasons for the probability estimates provided by the doctor were found to increase optimism versus negative reasons for the same event. They connect these findings to the evaluability work reported by Hsee and colleagues (1996), in that "an isolated numerical probability forecast is often difficult to evaluate and therefore does not have strong affective or intuitive implications" (Flugstad & Windschitl, 2003, pg. 108). This lack of evaluability is what leaves judgments based on the probability estimate open to the effect of scenario or other contextual information. This line of thought also leads to the notion that if particular judge's were better able to draw meaning from the numerical probability information, they would be less likely to be influenced by contextual information. This question will be addressed in the experiments below when individual differences in numerical ability are explored.

Hendrickx et al. (1989; 1992) also conducted several interesting experiments on the relation between scenario and probabilistic information. They presented subjects with descriptions of risky activities and asked them to decide whether to engage in the activity and to make ranked accident probability judgments. They manipulated the amount of supporting scenario information and whether frequency probability information was presented (e.g. "1 in every 25 experienced swimmers gets into trouble"). They found that more extensive concrete scenarios had a larger effect on perceived likelihood than abstract scenarios had. Additionally, they found that when scenario and frequency

information were presented together, the frequency information was dominated by the scenario information.

In summary, this research suggests that judgments based on precise numerical probability estimates can still be influenced by contextual information. A narrative summary of the relevant evidence supporting a probabilistic intelligence forecast may be a prototypical case of supplemental information affecting the interpretation of a numerical probability forecast.

Intuitive Judgments of Risk

In many forecasting situations, the forecaster is interested in communicating the risk associated with a particular event or activity, not just reporting the likelihood with which the event will occur. However, researchers have not reached a consensus on how risk should be defined or how risk is intuitively understood by the layperson. Brun (1994) discusses the many ways in which the risk concept has been defined. Although there are exceptions, Brun (1994) concludes that "...most definitions of risk include an estimate of uncertainty (a likelihood, possibility or judged subjective probability) for a negative event to happen (a possible loss or a negative consequence of an action). It follows that risk *perception* has a perceived probability/uncertainty aspect as well as a perceived severity aspect to it." (pg. 297). It follows, then, that many of the same issues and strategies that have been discussed in the context of probability estimation are also applicable to risk perception (Brun, 1994), including many of the theories discussed above. Interestingly, there are several ways in which the concept of risk may be different depending on the specific properties of the hazard that is under judgment. For example, Brun (1994) discusses research by Vlek and Stallen (1980) in which they state that risk may be "primarily associated with the probability of a loss whenever possible losses are small and of a similar magnitude and the probabilities are well specified, but that "risk" refers to the size of the loss (e.g. the possible magnitude or severity of an accident) in contexts where negative consequences can be serious, but the probabilities are vague and hard to assess." (pg. 297).

Further research has identified several other characteristics of the hazards that affect risk perceptions beyond some combination of likelihood and potential harm perceptions. Using a psychometric paradigm, in which laypeople are asked directly about their preferences and feelings toward different types of hazards, researchers have discovered several different factors, or underlying characteristics, that laypeople use to judge risk. The first factor is dread risk, which is defined by “perceived lack of control, dread, catastrophic potential, fatal consequences, and the inequitable distribution of risks and benefits”, while the second primary factor is unknown risk, which is defined as “unobservable, unknown, new, and delayed in their manifestation of harm.” (Slovic, 1987, pg. 283).

In addition to characteristics of the hazard, characteristics of the individual are likely to have strong effects on perceptions of risk. For example, recent research findings show that cultural outlooks and worldviews have a large impact on individuals’ feelings and perceptions of various societal risks (e.g. egalitarian versus individualistic worldviews). An individual’s worldview may have a stronger influence than other individual characteristics like race, education and political affiliation (Kahan, Braman, Slovic, Gastil & Cohen, 2007).

In summary, researchers generally agree that concepts of risk are composed of some combination of perceptions of likelihood and potential harm, although the layperson also uses characteristics of the particular hazard under judgment and personal worldviews in their perceptions of risk.

Consumer Perceptions of Probabilistic Forecasts and Risk Communications

Relatively few researchers have examined forecasts from the perspective of “consumers” (i.e. individuals using forecasts to make decisions) judging the quality or usefulness of forecasts (Fox & Irwin 1998; Yates, Price, Lee & Ramirez, 1996). One experimental paradigm comes from the business, law and meteorological domains, in which consumers are presented with the past predictive performance of a judge and then asked about the quality of the judge (e.g. Considering his/her past performance, which

judge would you like as your stock portfolio advisor?). In these experiments, the forecaster is providing single-event probability judgments for a series of cases, and the consumer is given multiple trials to learn about the past performance of each judge. Yates et al. (1996) and Price & Stone (2004), using methodologies as described above, found that consumers tended to prefer judges that were categorically correct (i.e. forecasted a probability of greater than .5 for events that occurred) and those that were more extreme to those that were better calibrated (see also Keren & Teigen, 2001). Calibration in this general sense refers to the extent to which a forecaster provides high probability estimates for events that do occur and low estimates for events that do not occur. Price and Stone called this latter effect the “confidence heuristic” and found that a more confident advisor (extreme in assigning probabilities) was thought to make more categorically correct judgments and was perceived to be more knowledgeable. Yates et al. (1996) also found evidence that consumers were sensitive to the reasons or explanations that accompanied the forecasts.

Keren & Teigen (2001) conducted a series of four experiments that suggest that lay people have a clear preference for more extreme and higher probabilities over less extreme ones (this is related to the “confidence heuristic” described above). They make a useful distinction in judging the “goodness” of probability judgments – namely, how informative is it (does it provide accurate information about the state of the world), and how valuable is it (is it useful for determining future actions to take). Subjects were given pairs of probabilities and asked which was more valuable and informative. The main finding was that the larger of the two probabilities was judged to be more valuable and informative.

In the second approach to studying consumer perceptions of forecasts and risk communications, consumers are presented with forecasts without additional frequency information about the past performance of the forecaster. Consumer trust and perceptions of source credibility have emerged as important factors in consumer perceptions of risk and overall perceptions of the quality or believability of risk forecasts (e.g. see McComas & Trumbo, 2001; Peters, Covelto & McCallum, 1997; Trumbo &

McComas, 2003). One interesting study reported by Trumbo & McComas (2003) examined how differences in perceptions of the credibility of government and industry reports of risk information related to how consumers process information, which leads to differences in perceptions of risk. Their results suggest that perceptions of low credibility promote systematic information processing, which leads to greater risk perceptions, whereas perceptions of high credibility for state or industry risk communication results in greater heuristic processing which leads to lower perceptions of risk.

Effects of the Format of Uncertainty Information on Perceptions of Risk and Perceptions of the Quality of the Forecast or Forecaster

Verbal versus Numerical Expressions of Uncertainty

There is quite a large literature concerning how people understand and use verbal statements to represent uncertainty (e.g. likely, seldom, very unlikely, etc; Budescu & Wallsten, 1995; Wallsten & Budescu, 1995). In general, verbal expressions of uncertainty have been found to be more vague than numerical estimates and, in some cases, have been found to relate to judgments that are less consistent and reliable than those based on numerical estimates (Wallsten, Budescu, and Zwick, 1993; see Fox & Irwin, 1998 for a discussion of the different research traditions relating to linguistic expressions of uncertainty). As noted above, however, numerical estimates of uncertainty are not immune to context effects or different interpretations by different consumers. In the case of communicating uncertainty information from forecaster to consumer, these research results suggest that using verbal statements of uncertainty is inferior to more precise numerical uncertainty estimates (Fischhoff, 2001; Heuer, 1999).

Fox and Irwin (1998) also review research that focuses on preferences for verbal versus numerical estimates of uncertainty. Overall, there is evidence to suggest that consumers tend to prefer to receive numerical uncertainty information but forecasters prefer to use verbal statements. In addition, Gurmankin, Baron, & Armstrong (2004) found that consumers were more trusting and comfortable with physician risk

information that included numerical probability estimates as opposed to verbal probabilities, although this effect interacted with the consumer's level of numerical ability (discussed below). Since the focus in this work is on consumers, these findings suggest that consumers of intelligence forecasts would prefer to have explicit numerical uncertainty information in forecasts.

Standard Probability Formats versus Frequency Formats

There has been a heated debate concerning the merits of frequency formats (i.e. 1 out of 10), as opposed to probability (i.e. .1) or percentage (10%) formats, as a more natural way of communicating uncertainty. Gigerenzer and colleagues (see Gigerenzer, 1994 for a review) have argued that humans are more naturally prepared to deal with frequency information, given that our species has evolved mechanisms to represent frequencies in order to learn from the natural environment. In fact, these authors have found that many of the standard biases identified in the heuristics and biases literature (e.g. overconfidence, base-rate fallacy and the conjunction fallacy) are not present when individuals are presented with frequency as opposed to standard probability information (Gigerenzer, 1994). The focus of the present investigation, however, is simply the transferal of risk/likelihood information from forecaster to consumer and does not involve statistical reasoning.

Additional research suggests that relative frequency information is easier to understand and is thought to be clearer than percentage or decimal representations of probability (see Burkell, 2004 for a discussion of this research). Overall, frequency representations of likelihood are thought to be more amenable to clear understanding and are easier to work with when additional calculations or comparisons need to be done to arrive at a judgment (Hoffrage, Lindsey, Hertwig and Gigerenzer, 2000; Burkell, 2004). However, Burkell (2004) mentions that “when the goal is only to present likelihood, and no statistical reasoning is required, percent format (e.g. 2%) is also appropriate [in addition to frequency formats], because it is perceived as easy to understand.” (pg. 204). In the risk communication situation modeled in the present work, this research suggests

that frequency or percentage formats may be effective in transferring information from forecaster to consumer.

Several researchers have also reported that consumer risk perceptions differ depending on whether probability is expressed as a relative frequency versus a decimal probability or percentage. Specifically, these results suggest that consumers perceive greater risk when presented with probability information as a relative frequency (Slovic, et al., 2000; Siegrist, 1997; Keller, Siegrist & Gutscher, 2006). For example, Slovic et al. (2000) found that consumers reported greater risk when the dangerousness of a mental patient was communicated as a relative frequency (e.g. “Of every 10 patients similar to Mr. Jones, 1 is estimated to commit an act of violence to others during the first several months after discharge”) as opposed to a percentage probability (e.g. “Patients similar to Mr. Jones are estimated to have a 10% probability of committing an act of violence to others during the first several months after discharge”).

In the present studies, frequency and percentage probability formats will be compared for the consistency with which consumers use this information (i.e. greater stated likelihood leads to greater perceived likelihood) and how they feel about intelligence forecasts with different formats for likelihood information.

Confidence Intervals and Reporting a Range of Plausible Probability Values in a Forecast

Several authors have suggested that some type of sensitivity analysis should accompany any probabilistic forecast. Presenting a probability point estimate in a forecast as well as a range of plausible values (i.e. a 95% confidence interval) gives the consumer information not only about the best probability estimate from the forecaster, but also relays information about the level of uncertainty inherent in the probabilistic analysis (sometimes called second-order uncertainty). The main goal of this approach is to present consumers with the most complete and honest information possible, with the hope that consumers will be able to use this information for further judgments and decisions.

Although including the results of uncertainty or sensitivity analysis in intelligence forecasts may seem beneficial, it is not clear that consumers will actually be able to use this information in judgment and decisions. In other words, there may be a tradeoff between more complete information and a consumer's ability to understand and use the information presented. Although not specific to presenting ranges of probability values, previous research suggests that more complete information can sometimes lead to a lack of understanding and inferior choices (Peters, Dieckmann, Dixon, et al., 2007; see Peters, in press for a brief review).

Hsee (1995) reports findings on the effects of presenting ranges of values on judgment and decision making. These findings suggest that when consumers are presented with a plausible range of values for an attribute, they may tend to ignore the information about this attribute and focus on other, possibly less relevant, information. In one study, Hsee (1995) showed participants two different files and asked them which one they would like to edit. One of the files was more interesting but paid less, while the other file was not as interesting but paid more. When the pay rate was presented to participants as a range, a larger percentage of participants chose to edit the more interesting but lower paying file. Thus, the range information appeared to allow the participants to weigh the pay rate less and focused them on the only other information on which to make the choice (i.e. how interesting the file was). This finding suggests that presenting numerical probability information as a range may cause consumers to ignore probability and focus on other information to make their judgments (e.g. the narrative summary of the evidence). Thus, one might expect consumers to rely more on the narrative information as opposed to the numerical probability information when the probability information is presented as a range.

There has also been research focused on how consumers feel about risk or likelihood information when it is presented as a range of plausible values. For example, in several studies, Johnson & Slovic (1995; 1998) presented participants with simulated risk communications from government sources like the Environmental Protection Agency (EPA). When these risk communications were reported with a range of plausible

probability values, participants tended to rate the agency as more trustworthy but less competent, and in verbal protocols many subject reported being uncomfortable with the range of probability values because it made them feel less confident that the EPA could estimate the risks involved (see also Johnson, 2003). In another example from a ~~completely~~ different domain, Epstein, Alper, & Quill (2004) reviewed the research literature relating to the presentation of clinical evidence to medical patients. They concluded that less educated and older patients did not like being presented with confidence intervals and had trouble understanding them.

External versus Internal Framing of Probabilistic Forecasts

There have been very few studies that have explored internal versus external framing of probability information in forecasts. An internal frame is thought to direct the consumer to interpret the probability estimate as a statement of uncertainty in the forecaster's personal belief or judgment (e.g. "I am 10% sure that x will happen in the next 6 months"), while an external mode may direct the consumer to interpret the probability as a statement about the propensity of the event in the external world, outside of the personal belief structure of the forecaster (e.g. "The probability that x will occur over the next 6 months is 10%").

In an unpublished manuscript, Fox & Malle (1997) discuss several interesting effects relating to internal versus external framing of probability information. Their results suggest that consumers tend to have more faith or belief in a forecaster that uses an internal frame for expressing probability as opposed to an external frame, and consumers feel that probability statements with an internal frame indicate that the forecaster is more certain and willing to take responsibility for the forecast (Fox & Irwin, 1998). One experimental result is particularly interesting in light of the consumer hindsight effects that will be the focus of Study 3. Fox and Malle (1997) presented experimental subjects with vignettes in which an economist forecasted that exports would increase in the next month. The numerical probability was framed as either internal (i.e. 70% sure) or as external (i.e. 70% chance) to the forecaster. Consumers where then told about the

eventual increase or decrease of exports over the next month. When told that exports increased in the next month, the majority of subjects said that they would rather promote the economist that reported his or her forecast in the internal frame, and, alternatively, if exports actually decreased, the majority of subjects said that they would rather fire the economist that reported his or her forecast in the internal frame (Fox & Irwin, 1998). In summary, consumers were more likely to praise a forecaster that they thought made a correct forecast and punish a forecaster that they thought made an incorrect forecast when the forecast was framed as internal to the forecaster.

Numeracy

The Conceptualization and Measurement of Numeracy

Numeracy defined in the broadest sense is the ability to understand and use numbers. This would include an understanding of the real number line, the ability to compare numbers in magnitude, the understanding of time and money, measurement, estimation, and the ability to perform simple arithmetic. At a somewhat higher level, a broad definition might also include basic logic, performing multi-step operations, a fundamental understanding of chance and basic statistical principles, and comfort with proportions, fractions, probabilities, and risks. Researchers have defined and measured numeracy in various ways, often because of differences in their specific research interests and domain of study. For example, Paulos (1988) defines innumeracy as the “inability to deal comfortably with the fundamental notions of number and chance” (pg. 3). He discusses difficulties individuals have in understanding extremely large and small numbers, grasping infinity, correctly using combinations and permutations to calculate quantities, and understanding basic concepts involving chance and probability. Another example comes from the healthcare domain, where researchers are often interested in the ability of the public to understand the risks and benefits of particular medical treatments. These authors often define numeracy as the ability to understand proportions, risks, percentages and probabilities, since these are the forms in which risk and benefit information is most often presented to consumers (Burkell, 2004).

Much of the quantitative health information presented to the public involves communicating the risks associated with particular diseases or treatment options (Burkell, 2004). Most of this information comes in the form of explicit probabilities, relative frequencies, and proportions, and it is assumed that people can interpret these different measures to make an assessment of the likelihood of different outcomes. Because of the importance and common use of this type of outcome likelihood information, researchers have developed numeracy measures specifically designed to assess these skills.

Schwartz, Woloshin, & Welch (1997) measured numeracy with three questions, which included a basic question assessing participants understanding of chance (i.e. how many heads would come up in 1000 tosses of a fair coin) and two questions asking the participants to convert a percentage to a proportion and a proportion to a percentage (i.e. the chance of winning a car is 1 in 1000, what is the percentage of winning tickets for the lottery?). This measure proved popular among researchers, and several authors have developed expanded versions of the original 3-item measure.

One important addition to the literature was the expanded numeracy measure created by Lipkus, Samsa & Rimer (2001). They added eight questions to the items from the Schwartz et al. numeracy scale. The additional items were designed to assess a participant's ability to understand and compare risks (e.g. Which of the following numbers represents the biggest risk of getting a disease: 1%, 10%, or 5%?) and to move between decimal representations, proportions and fractions. Peters, Dieckmann, Dixon, et al. (2007) have also used an expanded version of the Lipkus numeracy scale, introducing four more difficult items. Among other things, these additional items test the understanding of base rates as well as the ability to make more complex likelihood calculations.

Because the consumer of a probabilistic intelligence forecast is presented with very similar probabilistic estimates as consumers in the medical domain, numeracy defined as the ability to understand proportions, risks, percentages and probabilities is used in the present studies. Specifically, the expanded numeracy measure used by Peters et al.

(2007) is used in the studies reported below (see Appendix B for the expanded numeracy measure).

Previous Findings Relating to Individual Differences in Numeracy

Many researchers have tried to identify optimal methods of presenting numerical information to consumers (see Dieckmann, 2007). However, very little research has focused on how people with varying levels of numerical ability understand and use information presented in different formats. Fagerlin et al. (submitted) reviewed some of the literature on presenting risk and benefit information, but for the most part could only speculate about how individuals varying in numerical ability would deal with different presentation methods.

Peters, Vastfjall, Slovic, et al. (2006) conducted several experiments that examined how individuals varying in numerical ability were able to understand and use probabilities expressed in different formats and to what extent they were affected by information framing. In one study, Peters et al. (2006) examined whether numerical ability affected the perception of probability information. Participants were asked to rate the risk associated with releasing a hypothetical mental health patient. One half of the participants read the scenario in the frequency form (“Of every 100 patients similar to Mr. Jones, 10 are estimated to commit an act of violence to others during the first several months after discharge”) and the other half received the same information in percentage form (“Of every 100 patients similar to Mr. Jones, 10% are estimated to commit an act of violence to others during the first several months after discharge”). High numerate participants did not differ in their risk ratings between the two formats. Low numerate participants, however, rated Mr. Jones as being less of a risk when they were presented with the percentage format. The authors speculate that the low numerate, because of limited numerical skills, have more difficulty transforming one representation to another ($10/100 = 10\%$), and were therefore differently affected by the format. The low numerate may have reported a higher level of risk in response to the frequency format because in

this condition they generated more vivid images of the violent acts than in the percentage condition (Slovic, Finucane, Peters, & MacGregor, 2004).

In another study, researchers focused on trust and confidence in numerical information. Guarman, Baron, & Armstrong (2004) conducted a web survey in which they presented subjects with several hypothetical risk scenarios. The scenarios depicted a physician presenting an estimate of the risk that a patient had cancer in three different formats (verbal, numerical probability as percentage or numerical probability as fraction). Participants then rated their trust and comfort with the information, as well as whether they thought the physician distorted the level of risk. Numeracy was measured with a scale adapted from Lipkus et al. (2001). Overall, they found that participants were more trusting of the information in the numeric as compared to the verbal formats, although this effect interacted with numeracy. Even after adjusting for gender, age, and education, the results showed that those subjects with the lowest numeracy scores trusted the information in the verbal format more than the numeric, and those with the highest numeracy scores trusted the information in the numeric formats more than the verbal.

In summary, low numerate participants tend to be worse at reading survival graphs, more susceptible to framing effects, more sensitive to the formatting of probability and risk reduction information, and tend to trust verbal more than numerical information (Dieckmann, 2007).

Numeracy and Affective Processing

Peters and colleagues conducted two experiments that examined whether numerical ability was related to affective evaluations of numbers (Peters, et al., 2006). In one study, they used the jellybean task developed by Denes-Raj and Epstein (1994). Participants were presented with two hypothetical bowls of jellybeans and were told that they would win \$5 if they picked a colored jellybean. One bowl had a total of 100 jellybeans with 9 colored jellybeans. This bowl was labeled “9% colored jellybeans”. The second bowl had a total of 10 jellybeans with only one colored jellybean and was labeled “10% colored jellybeans”. Participants were then asked to choose the bowl they would like to pick from. They also rated how they felt about the 9% chance associated with the first

bowl (affect question), rated the precision of that feeling (“How clear a feeling do you have about the goodness or badness of Bowl A’s 9% chance of winning?”), and finally completed the expanded numeracy measure based on Lipkus et al. (2001). Participants lower in numeracy were more likely to choose from the bowl with the lower chance of winning (9% versus 10%) and they also reported less precise feelings about the 9% chance. The authors speculate that because the low numerate were not able to draw meaning from the percentage information, they were drawn to the objectively worse bowl by an irrelevant source of affective information – namely, the bowl with greater the number of winning beans.

In a second study, these authors used a task developed by Slovic, Finucane, Peters & MacGregor (2004). Two groups of participants are asked to rate the attractiveness of a simple gamble. The first group was given the following: 7/36 chances to win \$9 and 29/36 chances to win nothing. The second group was given a similar gamble but with a small loss: 7/36 chances to win \$9 and 29/36 chances to lose 5¢. The initial findings from Slovic et al. (2004) were that participants rated the gamble with the small loss considerably higher than the gamble with no loss. Peters et al. (2006) had participants complete this same task, but also had them complete measures of affect and affective precision, as well as the expanded version of the Lipkus et al. (2001) numeracy scale. They found that high numerate participants rated the bet with the small loss as more attractive than the bet with no loss, whereas low numerate participants rated the two bets as equally attractive. In this case the high numerate participants were actually making objectively worse judgments than the low numerate participants. The authors explain this difference by pointing out that the high numerate are actually better able to deal with numbers and therefore draw more affective meaning from numbers. In fact, high numerate participants were shown to have more positive and more precise feelings toward the 7/36 chances of winning, as well as more positive feelings toward the \$9. High numerate participants had particularly strong positive feelings toward the \$9 when it was accompanied by the small 5¢ loss, suggesting that they were particularly sensitive to the comparison of the small loss and the comparatively much larger gain. In this case, it

is possible the ability of the high numerate decision makers to draw meaning from numbers and number comparisons actually led them astray. As a whole, this work suggests that people lower in numeracy do not draw as much affective meaning from numbers, and consequently, may be more influenced by other, sometimes irrelevant sources of affective information.

In summary, previous work suggests that individuals lower in numeracy have difficulty judging risks and benefits, show larger framing effects, are sensitive to the formatting of probability information, trust verbal more than numerical information, and appear to not draw as much affective meaning from numbers. To my knowledge, individual differences in numerical ability have not been studied in the political forecasting domain, although there are many similarities between this domain and the experimental tasks that have been used in past research. Of particular relevance to the present studies is the sensitivity to the formatting of probability information, perceptions of narrative versus numerical information, and the finding that the low numerate may disregard any numerical probability information and focus on other sources of information when making judgments (e.g. narrative information in the forecast).

Hindsight Bias

In a series of experiments in the mid 1970's, Baruch Fischhoff (1975) demonstrated that people tended to overestimate the probability with which they would have correctly forecasted an event before it occurred when they were informed of the outcome of the event. In other words, once people know the outcome of an event they tend to overestimate how well they could have correctly predicted whether the event would happen or not without the outcome knowledge. For example, Fischhoff (1975) asked research participants to make a prediction about the outcome of a real world event, and then after the outcome of the event was known he asked them recall what they had predicted. On average, participants tended to be biased in the direction of the actual outcome of the event. Numerous follow-up studies have been conducted on a range of

related hindsight effects, and the hindsight literature has been reviewed by Hawkins and Hastie (1990).

Further explorations of the hindsight bias phenomenon showed that, as suspected, it was not just probability estimates that were biased in the direction of the actual outcome of the event, but relevant facts and evidence relating to the event were also reinterpreted in light of the outcome knowledge. In fact, it may be the causal reinterpretation of the evidence on which the forecast is based that is primarily responsible for hindsight effects (Hastie & Dawes, 2001). For example, Wasserman, Lempert & Hastie (1991) demonstrated that hindsight effects were only present when causal explanations relating the evidence to the outcome could be readily generated. In other words, when given knowledge about the outcome of the forecasted event, people naturally go back and try to make sense of the evidence in light of the outcome, using what Fischhoff (2001) calls a “heuristic of making sense” (pg. 544). “However, like other heuristics, rapidly integrating new information provides its benefits at a price. Those images of once-possible futures are no longer available when we need them. In their stead, we find pictures colored by our knowledge of what actually happened” (pg. 544).

Hindsight biases are a potential problem whenever consumers, or the forecasters themselves, revisit forecasts after the occurrence or non-occurrence of the forecasted event is known. Heuer (1999) discusses the potential problems that can occur with respect to hindsight bias in the intelligence forecasting domain. As discussed in the introduction, the auditing of forecasts is a necessary part of the learning process for forecasters, and forecast consumers also revisit forecasts, particularly after the occurrence of an event with negative consequences (i.e. the intelligence memo written about Bin Laden before September 11th, 2001). Fischhoff (2001) discusses the importance for forecasters to be precise in their forecasts, not only because this will help them learn from their mistakes, but also because ambiguous forecasts are more likely to result in hindsight bias on the part of future auditors. If the natural reaction is for people to take the relevant evidence and reinterpret it in light of the known outcome, then an ambiguous forecast may allow more opportunity for future auditors to see causal patterns in the set of

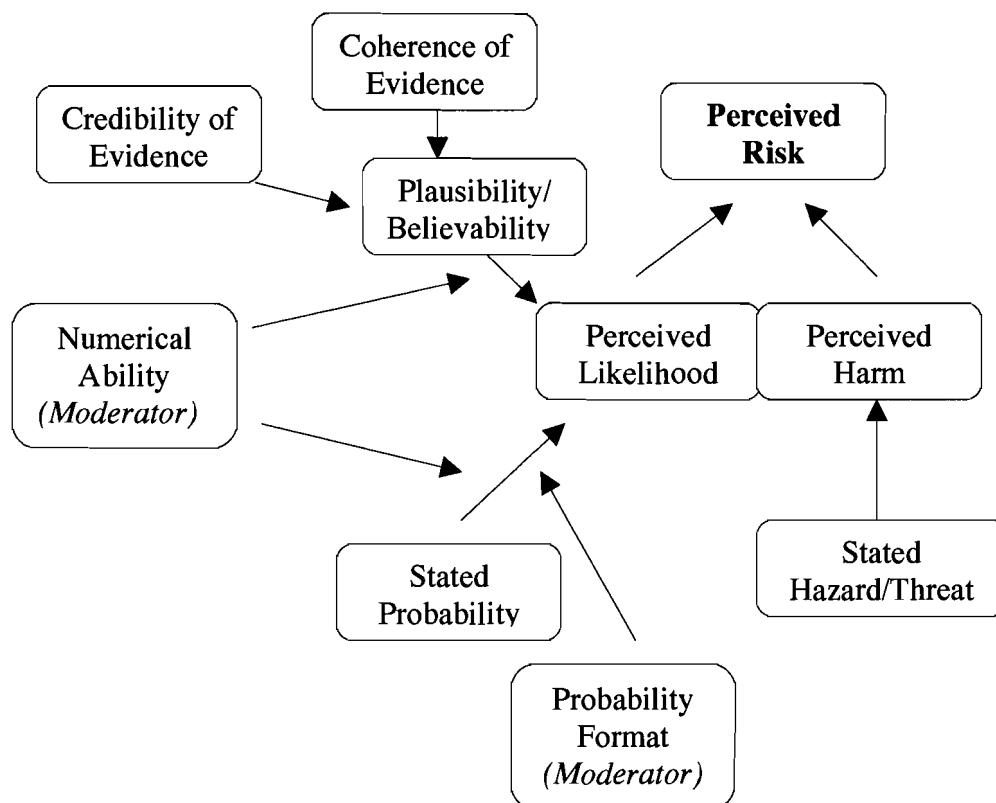
evidence that lead to the outcome. Fischhoff (2001) argues that more precise forecasts, preferably with numerical estimates of probability, are superior both in terms of forecaster learning and the reduction of the potential for gross hindsight reinterpretations of forecasts. These results suggest that in the intelligence forecasting tasks studied here, more ambiguous forecasts (pure narrative) would lead to larger hindsight effects than forecasts with probabilistic estimates (more precise). However, this hypothesis has so far not been tested.

CHAPTER III
A MODEL OF CONSUMER PERCEPTIONS OF SINGLE-EVENT INTELLIGENCE
FORECASTS AND PRIMARY RESEARCH QUESTIONS

A Model of Consumer Perceptions of Single-Event Intelligence Forecasts

Figure 1 shows a simple conceptual model of consumer perceptions of likelihood, potential harm and risk based on single-event intelligence forecasts that include both a discussion of narrative evidence and explicit estimates of likelihood and potential harm. This model is based on a conceptual analysis of the task of the intelligence consumer and the research literature reviewed in Chapter II. The properties of the intelligence forecast as well as the characteristics of the individual consumer are important in determining consumer perceptions of risk, as well as consumer feelings about the quality and usefulness of a forecast (e.g. source credibility, trust, competence, etc). Although the model depicted in Figure 1 is focused on consumer perceptions of likelihood, harm and risk, many of the same factors are expected to affect consumer feelings about the quality and usefulness of forecasts.

Figure 1. A simple model of consumer risk perceptions from simulated intelligence forecasts.



Risk perceptions are modeled as some unknown function of perceptions of the likelihood and the potential harm of the target event. The precise combination of likelihood and potential harm that compose risk perceptions is not known (and may vary depending on the particular characteristics of the task under study), and for this reason likelihood and potential harm are treated separately in the analyses presented below. In addition, global risk perceptions are also likely to depend on the characteristics of the particular hazard (e.g. the familiarity of the risk, or the amount of “dread” associated with the hazard) and the idiosyncratic perceptions of the individual consumer².

² Although the characteristics of the hazard and other idiosyncratic subject level effects are expected to affect individual perceptions of risk, they are not the focus of the present investigations and are not depicted in Figure 1.

Next, perceived potential harm is modeled as a function of the stated potential harm information presented in the forecast (e.g. statements about the expected loss of life or property if the target event were to occur). The perceived likelihood of the target event is a function of both the properties of the narrative discussion of the evidence set and the explicit probability provided by the forecaster. As consumers perceive greater coherence and credibility in the narrative evidence, they will perceive the target event as more plausible and believable and will therefore perceive the event to be more likely. In addition, as the explicit probability stated in the forecast increases, consumers will perceive the event to be more likely.

The relative reliance on the narrative evidence or the stated probability information is hypothesized to be moderated by the numerical ability of the consumer. Consumers that are higher in numerical ability will be able to evaluate and use the stated probability information, and will rely less on the narrative description for their perceptions of likelihood. Conversely, consumers lower in numerical ability will have more difficulty evaluating the probability information, and will therefore rely more on the narrative evidence for their perceptions of likelihood.

Finally, the format of the stated probability is expected to moderate the effect of stated probability on perceptions of likelihood. As discussed in Chapter II, some probability formats have been shown to be more easily evaluated by consumers, particularly consumers that vary in numerical ability.

Research Questions

The primary purpose of presenting explicit estimates of likelihood and potential harm is to communicate, as faithfully as possible, the estimates of risk that are generated by the analyst to the intelligence consumer. Using the model of consumer perceptions of intelligence forecasts developed above as a guide, there are several reasons why explicit risk estimates may fail in that goal: 1) consumers may focus on the vivid narrative information to such an extent that the probability information is neglected when judging risk (i.e. become overwhelmed by the vividness of the scenario information and

underweight the explicit probability information), 2) consumers lower in numeric ability may not understand the numerical information, and may just attend to the narrative information because they don't understand or want to avoid the numerical information, 3) even if consumers use the numerical probability information to some extent, they will not use it in the way that is intended (i.e. at least ordinal differentiation between probability values – 5% chance perceived as lower than 10% chance, etc), 4) consumer's perceptions of risk may not be consistent and may depend on the format of probability information, and 5) when judging a forecast in hindsight, the inclusion of numerical estimates in intelligence forecasts will affect consumer perceptions of the quality of the forecast. The overarching goal of this dissertation is to address, empirically, these five points concerning the inclusion of numerical probability and threat information in intelligence forecasts.

Below are the primary research questions that are the focus of this dissertation. Several specific hypotheses concerning specific manipulations (e.g. probability format) will be discussed in the context of each individual experiment.

1. To what extent does the presence of a narrative summary of the evidence supporting a forecast affect perceived risk and perceptions of the quality and usefulness of the forecast?
2. To what extent will perceptions of coherence in the narrative summary and credibility of the evidence affect perceived risk and perceptions of the quality and usefulness of the forecast?
3. Will consumers be sensitive to numerical information concerning probability and threat in the presence of a narrative summary of evidence, and will the majority of consumers be able to make at least ordinal differentiations between the probability levels? In addition, will consumers find intelligence forecasts with numerical probability and threat estimates more useful as decision making aids and also find them to be of higher quality?

4. Will the ability of consumers to use the numerical probability information in judging risk be moderated by the format of the probability information? Differences between verbal probability estimates, percentage formats, frequency formats, percentage formats presented as an external probability versus a confidence rating, and probabilities presented with a confidence range will be tested.
5. Will the numerical ability of the individual consumer affect the extent to which they rely on the numerical versus narrative information in a forecast? Also, will consumers with different levels of numeracy be able to use the probability information, and will they show different preferences for probability information in specific formats?
6. Finally, how will consumers perceive forecasts in hindsight (i.e. after they know the outcome of the forecasted event)? How will consumers perceive the quality of the forecast and, when they perceive a forecast to be “wrong”, will they place differing amounts of blame on the forecasters depending on whether numerical estimates of probability and threat were included in the forecast?

CHAPTER IV
THE SIMULATED INTELLIGENCE FORECASTS, THE SUBJECT POPULATION,
AND PRELIMINARY STUDY 1

The preceding chapters have detailed the conceptual task of the intelligence consumer and a model of the information sources and individual differences that affect perceptions of these forecasts. The development of the simulated intelligence forecasts and the subject population that was used in the primary experiments is discussed next.

Development of the Simulated Intelligence Forecasts

The simulated forecasts were modeled after a single-outcome indications and warning intelligence forecast, in which a narrative summary of pertinent evidence and explicit numerical information about the probability and potential harm of the event are presented. This type of forecast represents a relatively straightforward risk communication situation, with the assumption that if consumers have difficulty in this very simple case the problems would potentially be magnified in more complicated situations.

Four different forecasts were created that outlined the evidence relating to a potential terrorist attack in a large city in the United States (see Appendix A). Each scenario is approximately one-page long and was roughly modeled after historical examples of intelligence reports available in the public domain (e.g. see <http://www.foia.cia.gov/>). The now famous August 6th Presidential Daily Briefing (PDB) entitled “Bin Laden Determined to Strike in US” was also used as a rough template (see Appendix A). PDBs come in many forms but are generally relatively short intelligence products designed to alert and inform the president on matters of immediate import. The PDBs do not themselves include all of the information about how an analyst reached the conclusions that he or she did, although more information would be available in a separate more

detailed document or could be acquired through verbal questioning of the analyst, or other top managers and directors.

Each of the terrorist scenarios was approximately the same length with roughly the same quantity of evidence. The first scenario depicted a potential explosive attack against a government building in Washington, D.C., the second report depicted an explosive attack against a railway system in Chicago, the third report depicted an explosive attack against a passenger ship in a New York harbor, and the fourth report depicted a potential explosive attack against a professional sporting event in Los Angeles.

Experimental Participants

Ideally, real consumers of intelligence forecasts would have been recruited as study participants. For obvious reasons it is difficult to use actual intelligence consumers, since they include high-ranking government officials and advisors, members of congress, or the president. Several different populations of subjects were used in the present studies. A large community sample was used in Study II. A sample of graduate/law students from the University of Oregon was used in Study III (and Preliminary Study I), and a mixed sample of undergraduates from the University of Oregon and participants that had completed undergraduate degrees was used in Study IV.

Preliminary Study I: Pretesting the Simulated Intelligence Forecasts

Purpose

There were two primary goals of Preliminary Study I. The first goal was to assess the plausibility of the different explicit probabilities assigned to the simulated intelligence forecasts. In the experiments that follow, each scenario will be presented with several different explicit probability estimates. This study was designed to identify a plausible range of probability values that could be assigned to each scenario. It is important to make sure that after reading the simulated scenarios the participants, as a whole, were not completely surprised by the assigned probability. The proposed range of probability levels was between 1%-20%. This intermediate probability range was picked because the

probabilities were low enough to be believable in terms of a potential terrorist attack, yet they were still high enough to be, potentially, comprehensible for consumers. The goal was to test the hypotheses of interest in this range of probability and, in future work, investigate different ranges of probability. For instance, many potential terrorist attacks are very unlikely on the order of 1/1,000, or 1/1,000,000, and these small probabilities should be investigated in future work.

The second goal of this study was to explore consumer likelihood judgments concerning a target event when only a narrative description of evidence is available. It was expected that these likelihood judgments would be highly variable due to the idiosyncratic way in which consumers used the available evidence to make their judgments.

Procedure and Design

Participants read each simulated scenario and then rated whether they thought that each of the four possible assigned probabilities was a reasonable estimate of the likelihood of the target event. The assigned probabilities were 1%, 5%, 10% and 20%. Then they were asked to make their own estimates of the highest and lowest reasonable probability for the event. Each subject made ratings for all four scenarios. The order in which the subjects read the scenarios was counterbalanced, and the order in which they rated the probabilities was randomized.

Results and Discussion

A total of 17 psychology graduate students attending the University of Oregon participated in the study. Table 1 shows the percentage of participants who felt that the stated probability was a reasonable estimate, and Table 2 shows descriptive statistics for the subject-generated low and high probability ranges.

It is not clear from the percentages displayed in Table 1 that participants found the stated probabilities to be reasonable estimates of the chance that the terrorist attack would occur. For instance, in some cases only 55-60% of participants thought that a particular

probability was reasonable. Particularly concerning were the results for the 20% probability level for forecasts 3 and 4, which only a quarter of participants found acceptable. However, examination of the means and standard deviations for the high and low estimates for each forecast suggests that the stated probabilities are acceptable (see Table 2). For example, the mean low and high probability estimates for scenario 1 roughly span 1-20%. Taking the results as whole, it seems reasonable to attach stated probabilities ranging from 1% - 15% to forecasts 3 and 4, and stated probabilities ranging from 1%-20% for forecasts 1 and 2.

The second goal of this study was to explore the variance in perceived probability estimates. It is clear from the Table 2 that there is quite a bit of variability in perceived likelihood ranges between the subjects. For instance, some subjects reported likelihood ranges on the order of 20%-70%, while others reported likelihood ranges on the order of 0%-.01%. When only presented with narrative evidence, consumer perceptions of likelihood can vary widely. This demonstrates one of the disadvantages of using purely narrative reports, which tend to be ambiguous, to present risk information to consumers. Including explicit probability/risk information for consumers may help to alleviate this problem, and this hypothesis will be tested in the experimental work reported in the next chapter.

Table 1. Percentage of participants who felt that each probability was a reasonable estimate of the probability that the event would occur.

| | 1%? | 5%? | 10%? | 20%? |
|------------|------|------|------|------|
| Scenario 1 | 0.65 | 0.76 | 0.76 | 0.47 |
| Scenario 2 | 0.59 | 0.53 | 0.59 | 0.65 |
| Scenario 3 | 0.82 | 0.71 | 0.59 | 0.24 |
| Scenario 4 | 0.82 | 0.59 | 0.53 | 0.24 |

Table 2. Descriptive statistics for subject-generated probability ranges.

| | Low (%) | High (%) |
|-------------------|---------|----------|
| Scenario 1 | | |
| Mean | 1.54 | 21.71 |
| Median | 1 | 10 |
| SD | 2.03 | 19.53 |
| Extreme | 0 | 50 |
| Scenario 2 | | |
| Mean | 4.30 | 27.71 |
| Median | 1 | 30 |
| SD | 7.15 | 21.86 |
| Extreme | 0 | 70 |
| Scenario 3 | | |
| Mean | 1.13 | 14.36 |
| Median | 1 | 10 |
| SD | 2.34 | 15.01 |
| Extreme | 0 | 50 |
| Scenario 4 | | |
| Mean | 1.01 | 14.00 |
| Median | 0.1 | 10 |
| SD | 2.37 | 15.62 |
| Extreme | 0 | 50 |

CHAPTER V PRIMARY EXPERIMENTAL STUDIES

Overview of Primary Studies

Study 2. This study is an initial exploration of the effects of explicit estimates of likelihood and narrative scenario information on consumer risk perceptions. In addition, both the format of explicit likelihood information and the numerical ability of the consumers are explored as potential moderators of the effect of likelihood and scenario information on perceptions of risk.

Study 3. This study was designed to further test the impact of explicit likelihood information and specific properties of the narrative scenario information on perceptions of likelihood and potential harm. As in Study 2, the format of the explicit likelihood information and the numerical ability of the consumers are explored as moderators.

Study 4. The primary focus of Study 4 is to explore how consumers feel about intelligence forecasts in hindsight (with knowledge about the outcome of the forecasted event). Specifically, how do consumers feel about the usefulness and source credibility of these forecasts, and to what extent do they blame a forecaster when an event occurs that was given a relatively low likelihood in a forecast? Of particular interest are the types of information (i.e. narrative or stated likelihood) that consumers use to make usefulness, source credibility and blame judgments in hindsight.

Study 2 – Initial Explorations of the Effects of Explicit Likelihood and Scenario information on Perceptions of Risk

Purpose

The primary purpose of Study 2 was to determine whether consumers of simulated intelligence forecasts would be sensitive to stated likelihood information, particularly in

the presence of a narrative evidence summary. Another focus was whether consumers would be better able to use likelihood information in a particular format, and whether they would perceive particular likelihood formats to be higher in usefulness, knowledge and trust. In addition to assessing the information sources that consumers used to assess risk, it is also important to explore the types of forecasts that consumers are most comfortable using and feel are of the highest quality. Consumers are not likely to use forecasts that they have difficulty understanding or forecasts that they don't trust. Finally, the potential moderating influence of consumer numeracy was explored.

In the model of consumer risk perceptions developed above (see Figure 1), the properties of the scenario information and the explicit estimates of likelihood and potential harm are modeled as direct effects of perceived likelihood and perceived harm. Perceived likelihood and perceived harm then affect overall perceptions of risk. In this study, however, only global risk perceptions are measured. In Study 3, the effects of explicit likelihood, harm, and scenario information on perceptions of likelihood are studied directly.

Method

Participants

A community sample from the Eugene/Springfield area was recruited to participate in Study 2.

Procedure & Materials

Study participants were paid \$10 for approximately 1 hour of participation time. Each participant was asked to read one simulated intelligence report about a potential terrorist attack in Washington, D.C. (Scenario 1, discussed above). The report provided a narrative description of the evidence concerning the potential attack as well as a statement about the potential lives lost if the attack were to occur (see Appendix B). The statement about the potential lives lost was held constant for each participant. Since both stated probability and probability format were of prime interest in this experiment, stated

threat was held constant so as to minimize the effect that perceived threat had on risk judgments.

After reading the intelligence report participants then responded to a series of questions about what they read. In addition, participants completed the Lipkus numeracy scale and provided demographic information (see Appendix B). All study procedures were approved by the University of Oregon Institutional Review Board (IRB).

Experimental Design

The experiment was run as a fully between subjects 4 (uncertainty format) x 2 (probability) x 2 (evidence) design, with a total of 16 conditions. Uncertainty information was presented in four formats: verbal, frequency, percentage, and percentage w/range, and probability was presented as either highly unlikely (5%, 5 out of 100, Lo: 1% Best: 5% Hi: 10%) or fairly unlikely (20%, 20 out of 100, Lo: 10% Best estimate: 20% Hi: 30%). The verbal and numerical probability statements for each probability level were roughly matched based on previous research (Kent, 1994; Hamm, 1991), where highly unlikely was found to roughly correspond to a 5% probability and fairly unlikely was found to roughly correspond to a 20% probability. Bisantz, Marsiglio & Munch (2005) have used a similar approach in matching verbal and numerical probability statements. Additionally, the intelligence report was presented with either a summary statement only, or with a narrative description of the evidence and then the summary statement.

Dependent Variables

The dependent variables were designed to assess perceived risk and perceptions of the intelligence report. The primary measure of perceived risk was assessed with a single question: "How would you rate the risk associated with this possible attack?" Participants rated risk on a 0-10 scale ranging from "very low risk" to "very high risk".

In addition to perceived risk, participants also rated their perceived value or usefulness of the report: "How valuable is this intelligence report? In other words, does it

provide useful information for determining future actions to take?” In addition, participants rated how knowledgeable they thought the analyst was about this potential attack: “How knowledgeable does this analyst seem about this potential attack?” Both value and perceived knowledge were rated on 0-10 rating scales, ranging from “not at all valuable/knowledgeable” to “extremely valuable/knowledgeable. Finally, participants rated perceived trust in the summary and conclusions in the report: “How much do you trust that this analyst is giving you complete and unbiased information/conclusions about this potential attack?” Trust was rated on a 0-10 scale, anchored by “very little trust” and “very high trust”.

Results

Sample Characteristics

There was a total of n=305 participants (16-21 subjects per experimental condition) in a slightly unbalanced experimental design. Tables 3 and 4 show the sample characteristics.

Table 3. Sample Characteristics

| Characteristic | n | Mean (Median) | SD |
|------------------------------|-----|------------------|-------|
| Age | 305 | 48.31 (49.00) | 15.12 |
| Numeracy ^a (0-15) | 305 | 9.69 (10.00) | 3.17 |

^a Distribution is moderately negatively skewed.

Table 4. Sample Characteristics

| Characteristic | n | % |
|-------------------------------|-----|-------|
| Female | 182 | 59.7% |
| Education (n=304) | | |
| 8 th grade or less | 1 | .3% |
| Some HS | 15 | 4.9% |
| HS graduate | 85 | 27.9% |
| Vocational/trade school | 11 | 3.6% |
| Some college/2yr degree | 107 | 35.1% |
| 4yr college graduate | 39 | 12.8% |
| More than 4yr college degree | 46 | 15.1% |

Table 5 shows the Pearson correlations among the dependent variables in Study 1. Inspection of scatterplots for each variable pair confirmed that there were no non-linear associations between the variables. Thus, Pearson correlation coefficients were used as an appropriate index of the linear relationship between the variables.

Table 5. Pearson correlations (w/ 95% CI) between dependent variables (n=305).

| | Risk | Value | Knowledge | Trust |
|-----------|----------------------|----------------------|----------------------|-------|
| Risk | 1.00 | | | |
| Value | .358 (.256, .452) | 1.00 | | |
| Knowledge | .238 (.129, .341) | .649 (.579, .710) | 1.00 | |
| Trust | .144 (.032, .252) | .605 (.529, .672) | .728 (.671, .777) | 1.00 |

Perceived risk is moderately correlated with perceived value, and to a lesser extent with perceived knowledge and trust. Theoretically, one might expect roughly zero correlation between perceived risk and these variables. For example, regardless of

perceived risk a report may still be valuable or useful in terms of deciding on what to do about the risk. Because perceived risk is theoretically distinct from the other perception variables and the correlations are moderate, perceived risk will be analyzed in a univariate fashion.

Perceived value, knowledge, and trust were highly correlated in this sample. This probably reflects the fact that all of these items are getting at a similar construct of perceived “quality”. Due to the high correlations between these variables, it may be more parsimonious to treat them as indicators of a similar construct and analyze them together in multivariate analyses. See Appendix E for additional discussion of effect size measures, confidence intervals, and the statistical assumptions for analytic techniques used in Study 2.

Perceived Risk

The explicit likelihood estimates and the narrative discussion of the evidence were hypothesized to affect consumer risk perceptions. These effects may be moderated by the format of the likelihood estimates and the numerical ability of the consumers. Specific research questions are detailed below.

1. Will consumer perceptions of risk be affected by the explicit likelihood information presented in the forecast, in that higher stated likelihood will lead to higher perceptions of risk?
2. Will consumer perceptions of risk be affected by a narrative discussion of the evidence presented in the forecast, in that the presence of the narrative will lead to higher perceptions of risk?
3. Will the format of the likelihood information moderate the effect of explicit likelihood on perceived risk?
 - a. The sensitivity of consumers to the different levels of stated likelihood may be moderated by the format of the likelihood information.

- b. In addition, previous research suggests that expressing likelihood in a frequency format results in higher estimates of risk than equivalent expressions of likelihood in decimal or percentage form.
4. Will consumers lower in numeracy have more difficulty using the explicit probability information to inform their risk judgments, and will their judgments be moderated by the format of explicit likelihood?

Tables 6 and 7 show the effects of explicit likelihood information and uncertainty format on perceptions of risk both with and without a narrative description of the evidence.

Table 6. The effect of explicit likelihood estimates and uncertainty format on risk perceptions without a narrative description of evidence.

| | Verbal | Frequency | Percentage | Percentage w/range | Total |
|----------------------------------|---------------------|---------------------|---------------------|-----------------------|---------------------|
| Highly Unlikely (5%, 5/100) | 4.55 (2.44) n=20 | 4.05 (1.99) n=21 | 3.85 (2.48) n=20 | 3.82 (1.70) n=17 | 4.08 (2.17) n=78 |
| Fairly Unlikely (20%, 20/100) | 5.40 (1.60) n=20 | 5.17 (2.46) n=18 | 4.75 (1.97) n=20 | 4.82 (2.43) n=17 | 5.04 (2.10) n=75 |
| Total | 4.98 (2.08) n=40 | 4.56 (2.26) n=39 | 4.30 (2.26) n=40 | 4.32 (2.13) n=34 | |

Note: Mean (SD) and sample size (n) are reported.

Table 7. The effect of explicit likelihood estimates and uncertainty format on risk perceptions with a narrative description of evidence.

| | Verbal | Frequency | Percentage | Percentage w/range | Total |
|----------------------------------|---------------------|---------------------|---------------------|-----------------------|---------------------|
| Highly Unlikely (5%, 5/100) | 5.67 (2.35) n=18 | 5.24 (2.36) n=21 | 4.20 (2.73) n=20 | 4.89 (1.91) n=18 | 4.99 (2.38) n=77 |
| Fairly Unlikely (20%, 20/100) | 5.05 (1.85) n=20 | 5.32 (2.81) n=19 | 5.40 (2.48) n=20 | 5.38 (2.85) n=17 | 5.28 (2.46) n=75 |
| Total | 5.34 (2.10) n=38 | 5.28 (2.55) n=40 | 4.80 (2.64) n=40 | 5.12 (2.38) n=34 | |

Note: Mean (SD) and sample size (n) are reported.

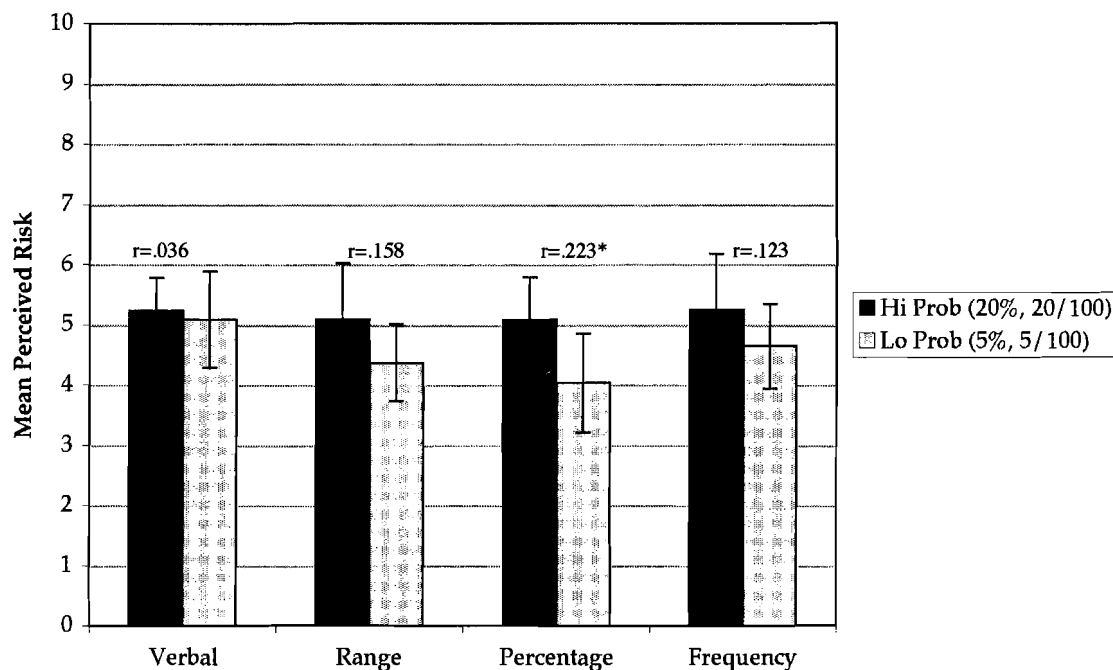
Consistent with the first two hypotheses, both higher stated probability, $F(1,299) = 5.74$, $p = .017$, $r = .136$ (95% CI = .245, .024), and the presence of a narrative summary of the evidence, $F(1,299) = 4.99$, $p = .026$, $r = .126$ (95% CI = .235, .014), resulted in higher perceptions of risk³. In addition, the frequency format elicited slightly higher risk perceptions than the percentage format, $t(299) = 1.08$, $p = .28$, $r = .062$ (95% CI = .173, -.051), although this effect was small and not statistically significant.

Consumers, averaging across the format condition, were sensitive to the stated likelihood information when judging the risk of the potential terrorist attack. Simple effects were used to test whether this difference in perceived risk was present for each likelihood format. Figure 2 shows the mean perceived risk between the two levels of stated probability for each probability format. Effect sizes are also presented for the difference between the stated probability conditions for each format (* indicates a contrast is significant at $p < .05$). Consumers did not show substantially different perceptions of risk when likelihood was expressed in a verbal form. However, risk perceptions did differ in the expected direction in each of the numerical likelihood formats, although only the percentage format elicited a significant effect. It is also

³ See Appendix E for a discussion of the General Linear Model (GLM) used to model consumer perceptions of risk, as well as a discussion of the r effect size measure.

interesting to note that there were virtually no differences between the probability formats in terms of perceived risk at the higher probability value, while at the lower probability level there were substantial differences between the formats.

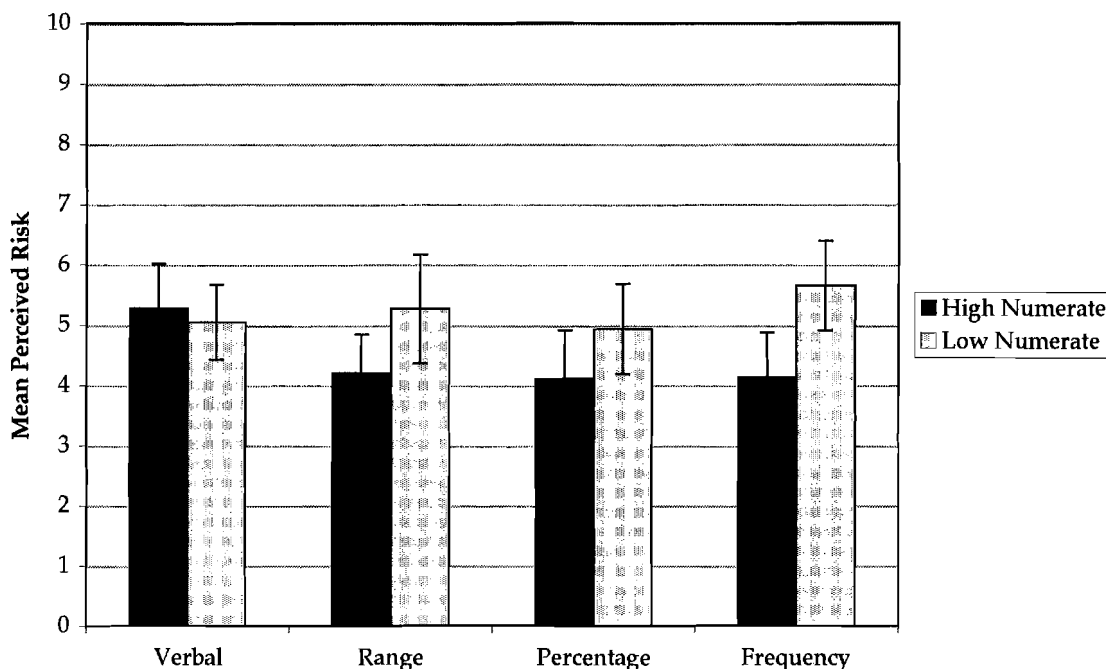
Figure 2. The effect of stated likelihood and likelihood format on perceived risk.



Note: * indicates that the contrast was significant at $p < .05$.

To explore the last set of hypotheses, individual differences in numeracy were used to predict perceived risk. There was a small to medium effect of numeracy, $F(1, 298) = 6.47$, $p = .011$, $r = .148$ (95% CI = .256, .036), such that consumers lower in numeracy reported higher perceptions of risk. In addition, numeracy level interacted with the format of the likelihood information to affect perceptions of risk. Figure 3 shows the mean perceived risk for the format conditions split by high and low numeracy. A median split for numeracy is used for simplicity of display.

Figure 3. The effect of format condition and numeracy level on perceptions of risk.



There was a significant interaction between numeracy and the contrast between the verbal probability condition and the numerical conditions combined, $t(295) = 2.27$, $p = .029$. The low numerate showed little difference between the verbal probability condition and the numerical conditions ($r = -.05$, 95% CI = .106, -.204) while the high numerate showed higher perceived risk in the verbal condition and decreased perceptions of risk in the numerical conditions ($r = .252$, 95% CI = .398, .093). The high numerate were sensitive to the numerical probability information and showed a decrease in perceived risk compared to the verbal, while the low numerate perceived roughly the same level of risk in the verbal condition as compared to the numerical conditions. However, the low numerate do show a trend toward higher average risk ratings in the frequency condition as compared to the other conditions ($r = .124$, 95% CI = .274, -.032), which is consistent with previous findings reviewed above.

Since consumers lower in numeracy have more difficulty evaluating numbers, it follows that the explicit likelihood information may not affect their perceptions of risk in

a consistent manner (i.e. higher perceptions of risk with higher stated likelihood regardless of the format of the likelihood information).

Figures 4-7 show the relationship between uncertainty format and stated likelihood by numeracy (median split for display purposes) and evidence condition. Note that due to the sample size there are only 8-12 participants included in each of the means displayed in these graphs. Statistical power is a definite concern when testing the simple effects for these subgroups. Effect size measures are presented in the figures.

Figure 4. The effect of stated likelihood and likelihood format on perceived risk with summary only, for consumers low in numeracy.

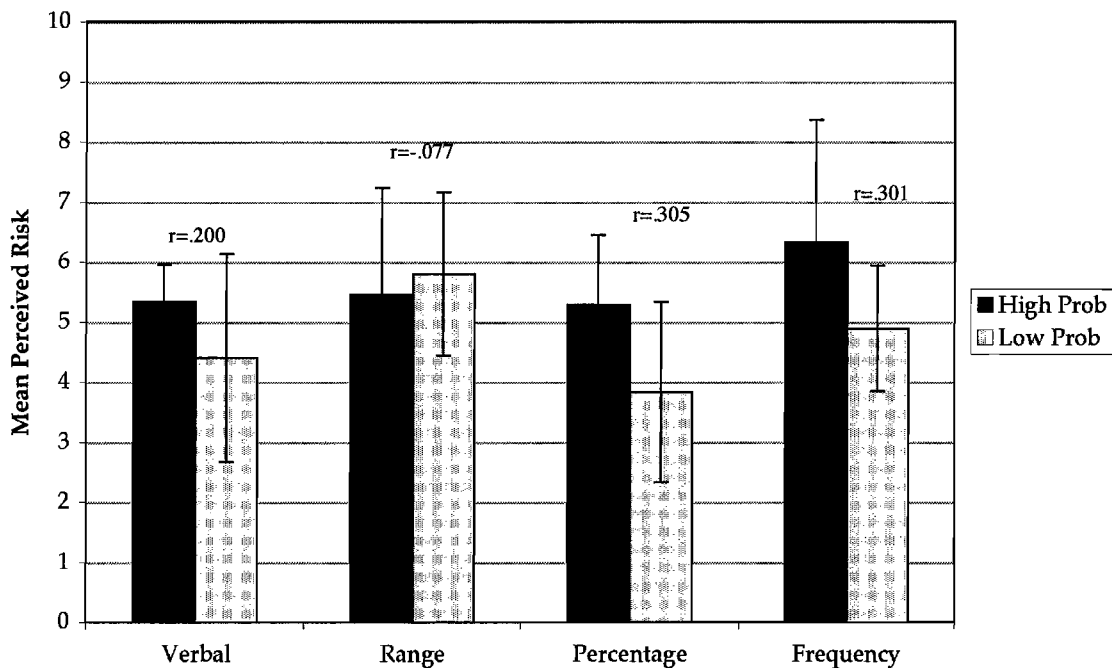


Figure 5. The effect of stated likelihood and likelihood format on perceived risk with summary only, for consumers high in numeracy.

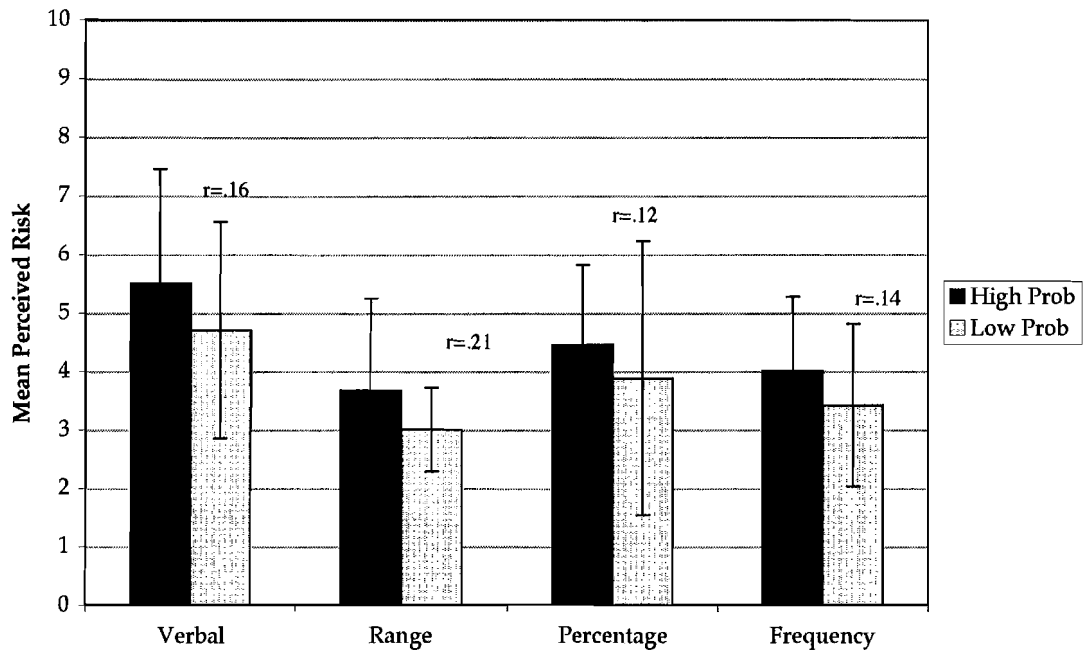


Figure 6. The effect of stated likelihood and likelihood format on perceived risk with summary plus evidence, for consumers low in numeracy.

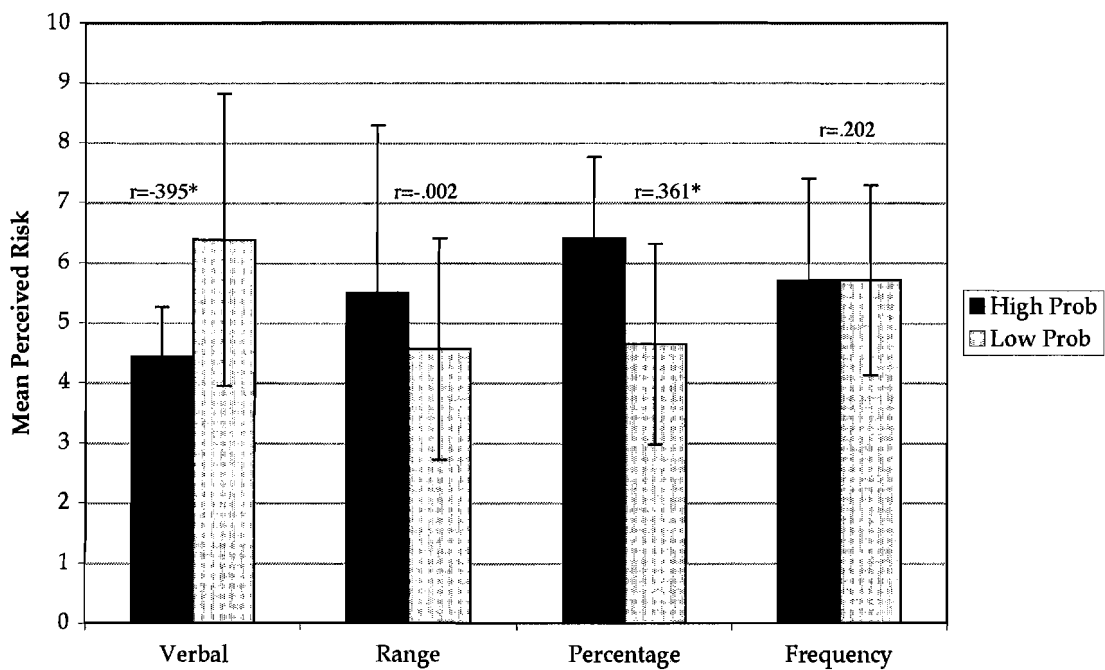
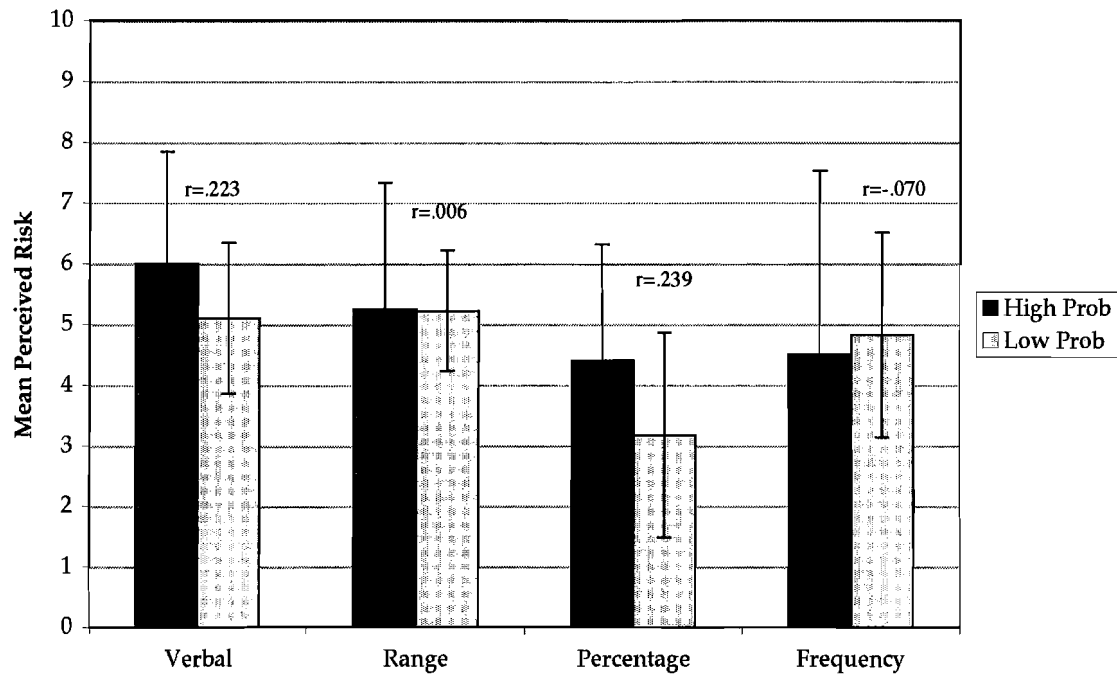


Figure 7. The effect of stated likelihood and likelihood format on perceived risk with summary plus evidence for consumers high in numeracy.



In the summary only condition, the low numerate show the expected pattern of higher risk ratings in the higher probability condition under all formats except for the percentage w/range format, while the high numerate show the expected pattern in all formats. When the evidence is present as well, however, the patterns change for both the high and low numerate. The low numerate show the expected pattern for the percentage and percentage w/range conditions, show no differentiation in the frequency format, and show a large effect in the opposite direction in the verbal condition. The high numerate show a relatively large effect in the expected direction with the percentage format, and show a flat trend or slightly opposite effect in the frequency and percentage w/range formats. To summarize, only the percentage format showed the predicted relationship at both levels of the evidence condition for both the high and the low numerate participants, although these effects clearly need to be replicated in a more powerful experimental design. These results also suggest that the presence of the narrative evidence summary has a strong effect on how consumers use the stated likelihood information to inform

their risk judgments. The competition between the narrative evidence summary and the stated likelihood information will be examined in more detail in Study 3.

Perceived Usefulness of the Forecast and Perceptions of Knowledge and Trust

Perceptions of the forecast were explored as a function of the format of explicit likelihood information, the presence of the narrative description and the numerical ability of the consumer. Specific research questions are outlined below:

1. Will the presence of a narrative discussion of the evidence affect consumer perceptions of the usefulness of the forecast and/or perceived knowledge and trust?
2. Will the format of the likelihood information affect consumer perceptions of the usefulness of a forecast and/or perceived knowledge and trust?
 - a. Because of the difficulty in interpreting verbal probability estimates, verbal probability statements will be perceived as less useful, and the forecaster will be perceived as less knowledgeable and trustworthy.
 - b. Previous research also suggests that consumers may perceive a forecaster that presents a probability point estimate with a range as less knowledgeable than one that presents a point estimate only. Point estimates with a range may also be perceived as less useful because there is not a single number on which a consumer can use to help assess the risk of the target event.
3. Will consumers varying in numeracy prefer likelihood information in different formats?
 - a. Previous research suggests that the low and the high numerate may prefer probability information in different formats. The low numerate will perceive greater usefulness, knowledge and trust in forecasts with frequency representations of likelihood as compared to percentage representations.

Because of the high intercorrelations between the perception variables, a multivariate general linear modeling framework was used to assess the effects of the independent variables on the linear combination of the three perception variables.

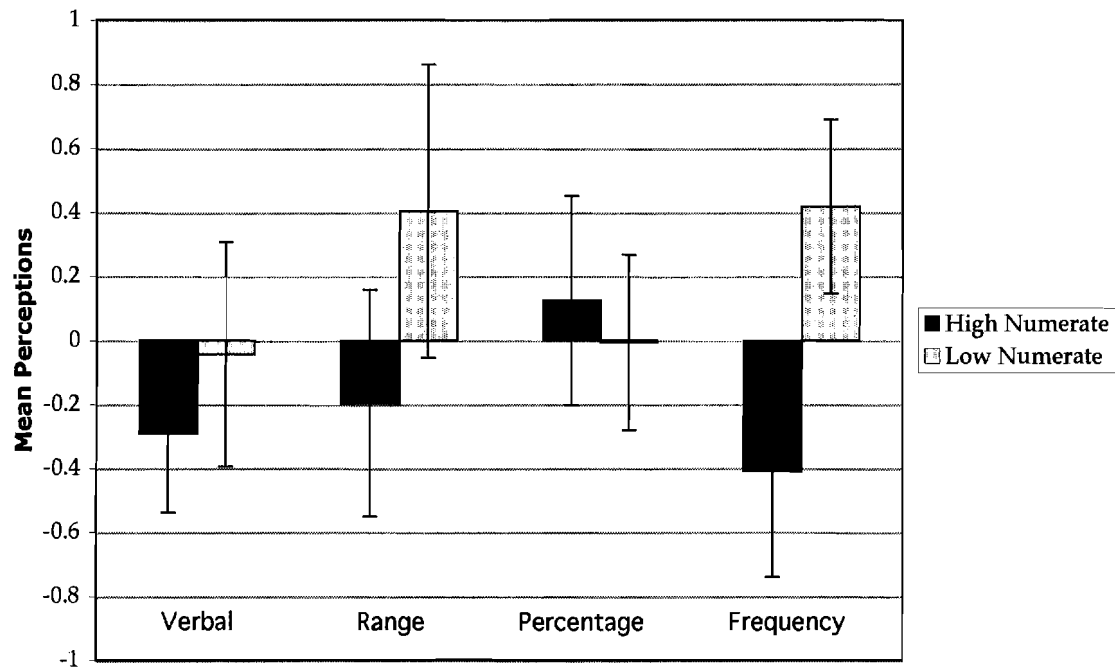
As hypothesized, participants who read a narrative summary of the evidence with the forecast reported higher levels of perceived usefulness, knowledge and trust, Pillai's = .053, $F(3,296) = 5.49$, $p = .001$, $r = .230$ (95% CI = .334, .121)⁴. The standardized discriminant function coefficients for the linear combination of the perception variables were -.709, -.509 and .145 for usefulness, knowledge and trust, respectively. This indicates that all of the perception variables contributed to the differentiation of the evidence groups, although perceived usefulness made the largest contribution.

To address the second set of hypotheses, differences in consumer perceptions among the different likelihood formats were explored next. The first contrast compared the verbal probability condition to the average of the numerical conditions, and although it was not statistically significant, Pillai's = .013, $F(3,296) = 1.31$, $p = .271$, $r = .11$ (95% CI = .220, -.002) the effect was in the expected direction, with lower ratings of value, knowledge and trust in the verbal likelihood condition. However, there was virtually no difference between the percentage with range condition and the average of the other numerical likelihood conditions, Pillai's = .001, $F(3,296) = .11$, $p = .957$, $r = .031$ (95% CI = .143, -.082). This is inconsistent with previous findings in which forecasters who presented likelihood estimates with ranges were perceived as less knowledgeable than those who presented likelihood point estimates only.

The final set of hypotheses concern the potentially moderating influence of consumer numeracy. Figure 8 shows the effect of uncertainty format and consumer numeracy on perceptions of usefulness, knowledge and trust.

⁴ Details of the multivariate analyses are discussed in Appendix E.

Figure 8. The effect of uncertainty format and consumer numeracy on perceptions of usefulness, knowledge and trust.



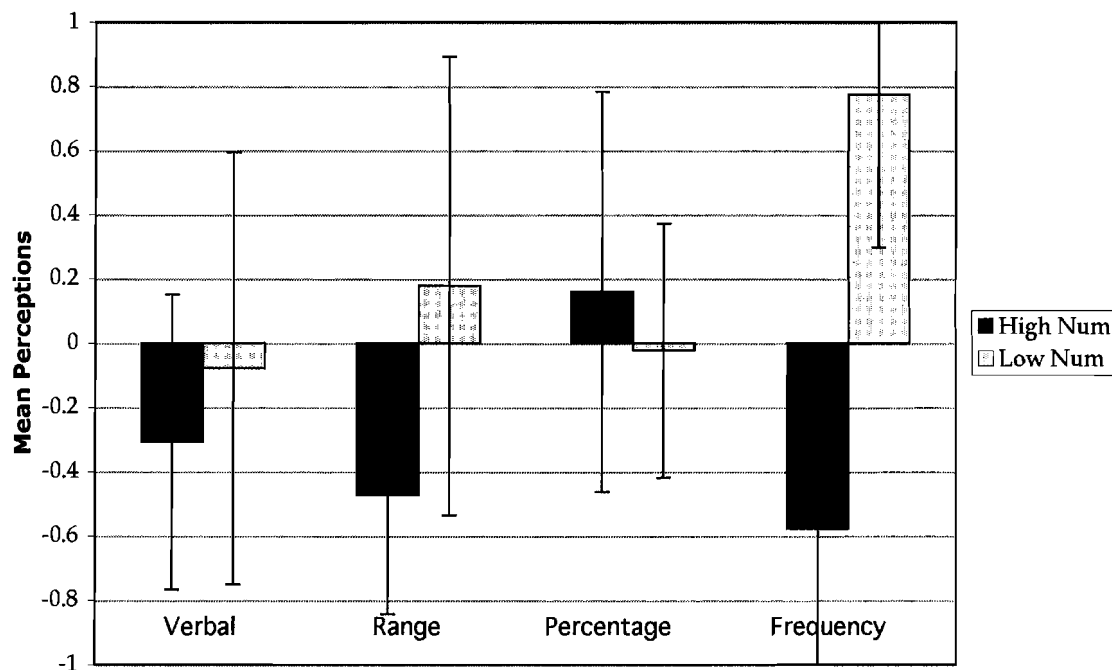
Note: The dependent variable in this figure is the linear combination of perceived usefulness, knowledge and trust used in the MANOVA.

As a main effect, participants lower in numeracy reported greater perceptions of usefulness, knowledge and trust, Pillai's = .07, $F(3,296) = 7.16$, $p < .001$, $r = .26$ (95% CI = .362, .152). The standardized discriminant function coefficients were -.442, -.524, and -.166 for value, knowledge and trust, respectively. Furthermore, as hypothesized, the difference in perception ratings between the frequency format and the percentage format was moderated by the numeracy level of the consumer, Pillai's = .029, $F(3,294) = 2.89$, $p = .036$, $r = .170$ (95% CI = .277, .059), and perceived trust was the main variable driving this effect (standardized discriminant function coefficients were -.416, .158, and -.818 for value, knowledge, and trust, respectively). The low numerate rated the frequency condition higher than the percentage condition, and the opposite was true for the high numerate. Overall, the high numerate found a forecast with a percentage likelihood estimate to be more useful, and higher in knowledge and trust than a forecast with the other uncertainty formats. The low numerate found a forecast with a frequency and

percentage with range likelihood estimate to be more useful, and higher in knowledge and trust than a forecast with a percentage or verbal likelihood estimate.

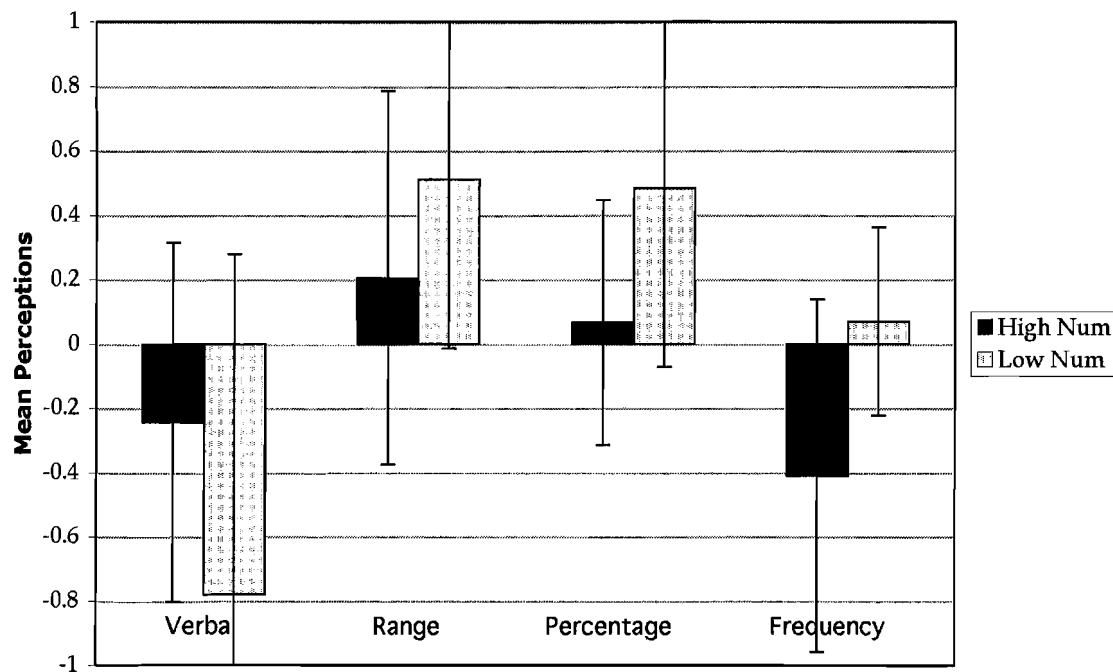
However, additional exploratory analysis revealed that the preference for the percentage format by the high numerate and the frequency format for the low numerate were moderated by stated likelihood, Pillai's = .04, $F(3,287) = 3.58$, $p = .014$, $r = .20$ (95% CI = .305, .090). Figures 9 and 10 show the effect of uncertainty format and numeracy at each level of stated likelihood. The figures show that the preference for the frequency format by the low numerate and the preference for the percentage format by the high numerate is only present at the lower level of likelihood.

Figure 9. The effect of uncertainty format and numeracy on perceptions of usefulness, knowledge and trust at low stated likelihood (i.e. 5%, 5/100).



Note: The dependent variable in this figure is the linear combination of perceived usefulness, knowledge and trust used in the MANOVA.

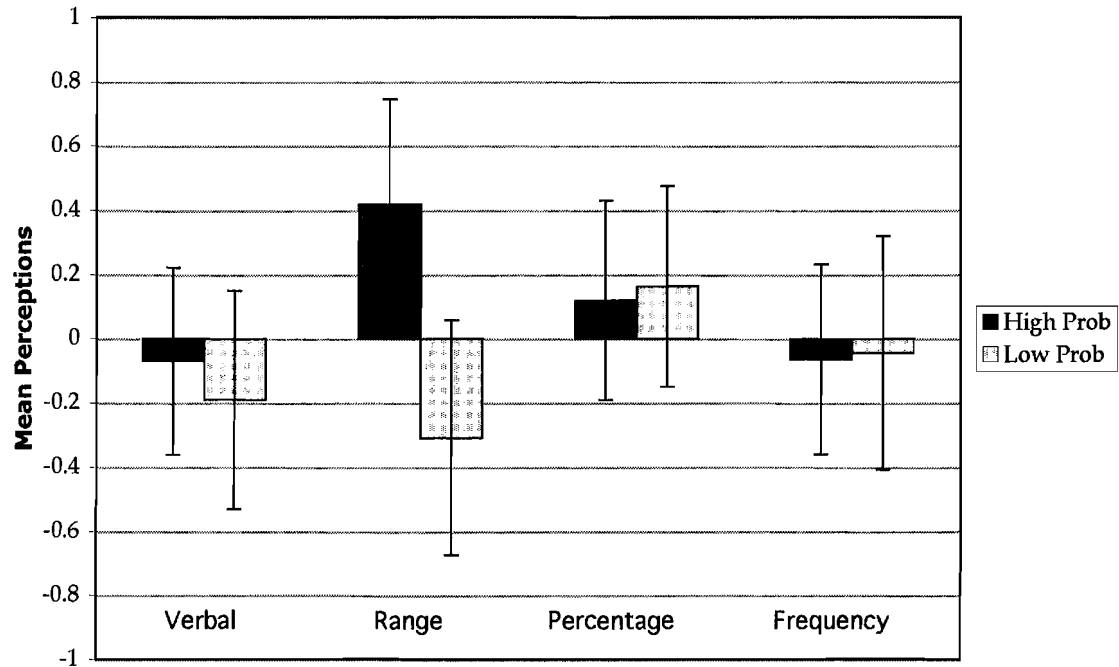
Figure 10. The effect of uncertainty format and numeracy on perceptions of usefulness, knowledge and trust at high stated likelihood (i.e. 20%, 20/100).



Note: The dependent variable in this figure is the linear combination of perceived usefulness, knowledge and trust used in the MANOVA.

In addition, collapsing across numeracy, stated likelihood moderated the effect of uncertainty format on perceptions of usefulness, knowledge and trust. Figure 11 shows the effect of uncertainty format and stated likelihood on perceptions of usefulness, knowledge and trust. Perception ratings were roughly equal for each likelihood format except for the percentage with range condition. The percentage with range condition was rated substantially higher in the high probability condition than in the low probability condition, Pillai's = .03, $F(3,298) = 2.95$, $p = .033$, $r = .341$ (95% CI = .536, .112), and perceived usefulness was the main variable driving this effect (standardized discriminant function coefficients were -.822, -.502, and .332 for value, knowledge and trust, respectively). It appears that consumers only found the percentage with range condition useful in the high probability condition, and it is possible that when the probability values get too low consumers can no longer use the range information.

Figure 11. The effect of uncertainty format and stated likelihood on perceptions of usefulness, knowledge and trust.



Note: The dependent variable in this figure is the linear combination of perceived usefulness, knowledge and trust used in the MANOVA.

Summary and Discussion

The primary purpose of Study 2 was to determine whether consumers of simulated intelligence forecasts would be sensitive to stated likelihood information, particularly in the presence of a narrative evidence summary. Another primary focus was whether consumers would be better able to use likelihood information in a particular format, and whether they would perceive particular likelihood formats to be higher in usefulness, knowledge and trust. Finally, the potential moderating influence of consumer numeracy was explored.

The Effect of Stated Likelihood and Narrative Information

On average, the magnitude of the stated likelihood had an effect on perceived risk, suggesting that participants were sensitive to the explicit likelihood information in the forecast. The presence of a narrative summary of the evidence also resulted in increased perceptions of risk. Presumably, the narrative information formed a more compelling story and elicited more compelling imagery from consumers. The compelling story likely made the attack seem more plausible, which led to the higher ratings of risk. The effect sizes were roughly equal for the explicit likelihood and narrative conditions suggesting that including a description of the evidence underlying a forecast has roughly the same effect on perceived risk as a stated probability shift from 5%-20%. In addition, not only did the narrative information increase risk ratings, but consumers also found the forecast to be more useful and the forecaster more knowledgeable and trustworthy.

The Effects of Likelihood Format

The results so far have provided evidence that consumers are sensitive to the probability information in the forecast. As expected, however, the extent to which consumers were sensitive to the stated likelihood was moderated by the format of the likelihood information.

Verbal Likelihood Format. Because of the lack of specificity of verbal probability statements, consumers were not, as a whole, sensitive to changes in verbal stated likelihood. Consumers also tended to rate forecasts with verbal estimates of likelihood lower in terms of usefulness, knowledge, and trust compared to the numerical formats. This result adds to the long list of indictments against verbal probability statements. It is clear from these findings and previous research discussed in Chapter II that verbal probability statements without a reference scale are not particularly helpful in transmitting risk information from analyst to consumer.

Numerical Likelihood Formats. In general, consumers were sensitive to the changes in explicit likelihood in each numerical condition, although the percentage format showed the largest effect. One primary effect of interest was whether both high and low

numerate participants were sensitive to the stated likelihood information (higher perceived risk in the higher likelihood condition) for each of the likelihood formats. Inspection of the mean perceived risk for each condition revealed inconsistent risk ratings in several conditions (see discussion above). Consumers appeared to be particularly insensitive to the stated likelihood information when they read a narrative about the evidence, presumably because they were not focusing on the likelihood information as much in the presence of the narrative summary. Without the narrative summary, however, the low and the high numerate showed roughly consistent patterns of risk perception across stated likelihood for each numerical likelihood format (except for the range condition for the low numerate). In the end, however, only the percentage likelihood format consistently showed the expected pattern across the likelihood levels for both high and low numerate consumers with and without a narrative summary of the evidence.

Frequency versus Percentage Likelihood Formats. Consumers also reacted differently to the likelihood formats based on stated likelihood and numerical ability. The low numerate perceived greater usefulness, knowledge, and trust in the frequency condition as compared to the percentage condition, and the opposite was true for the high numerate. This pattern of results, however, was only present at lower stated probability, possibly because people generally have more difficulty dealing with these probabilities and are therefore more sensitive to format. In addition, it is noteworthy that although the low numerate expressed higher ratings of usefulness, knowledge and trust for likelihood information in frequency form, they did not, on average, consistently perceive higher risk when presented with higher stated likelihood in frequency form.

Percentage with Range Format. Consumers rated the percentage w/range likelihood format higher in usefulness, knowledge and trust when the forecast involved higher likelihood values (i.e. best estimate 20%) as compared to lower likelihood values (i.e. best estimate 5%). Considering the difficulty that many people have understanding low probabilities, ranges of plausible likelihood values in the low probability range may not be useful for consumers.

In addition, risk perceptions in the percentage w/range condition were not as consistent (i.e. consistency would be higher risk perceptions for higher stated likelihood ranges) as risk perceptions in the percentage format. Although the range format actually provides more information than the other numerical formats (i.e. high, low, and best estimates of probability), the additional information may make it difficult to use the estimates to inform perceived risk. For example, a particular consumer could be a pessimist and focus on the high end of the range when judging risk or could focus on the low or best estimate depending on his or her inclination. In many ways the range condition is superior to the point estimate formats, in that the consumer is also given an idea of the certainty that the forecaster has in the estimate. Because of the present interest in the risk communication and forecasting literature on sensitivity analysis and reporting ranges of parameters, I focus on the percentage with range condition again in Studies 3 and 4.

Study 3 – Further Investigations of Including Numerical Estimates of Likelihood and Harm in Forecasts

Purpose

Study 3 was designed to address two primary issues. The first issue was the direct comparison of purely narrative intelligence forecasts to forecasts with narrative as well as numerical estimates of likelihood and potential harm. Although people recommending probabilistic analyses often assume that providing numerical estimates is superior to purely narrative forecasting (e.g. Fischhoff, 2001), this has not been shown empirically in the intelligence domain. Providing numerical estimates of likelihood, for instance, should facilitate the communication of probability information from analyst to consumer in a more accurate, consistent manner than verbal probability statements or purely narrative descriptions of evidence. However, as was discussed in Chapter II, numerical estimates of likelihood can still be affected by contextual factors (e.g. a narrative evidence summary accompanying a quantitative forecast), and may not be consistently

interpreted by different consumers. Thus, it is important to empirically explore the effects of including quantitative estimates in intelligence reports as compared to purely narrative reports. In Study 2, the verbal condition included verbal probability statements that were used in the place of quantitative estimates. The use of verbal labels in place of quantitative probability estimates is an interesting issue in its own right, but the more fundamental question is whether purely narrative reports (with no quantitative or verbal probability summary) are different from reports with quantitative estimates.

The second issue was to more directly assess consumer's use of narrative and explicit likelihood information when forecasts include both of these information sources. Specifically, the goal was to determine the impact of explicit likelihood information and specific properties of the narrative on perceptions of likelihood. Unlike Study 2, in which perceived risk was the primary dependent measure, the impact of explicit likelihood and narrative information is related directly to perceptions of likelihood. Perceived likelihood and perceived potential harm are thought to be two of the sources of information that affect global perceptions of risk (see Model of Consumer Perceptions of Forecasts in Chapter III).

Additionally, as in Study 2, both the numerical ability of the consumers and the format of explicit likelihood information were explored as potential moderators (see Model of Consumer Perceptions of Forecasts in Chapter III). Based on experimental results reported by Fox and Malle (1997), in which the internal or external framing of a subjective probability estimate affected consumers perceptions of the forecaster, internal and external framing of likelihood were explored as a potential moderator variable. In addition to these point estimate likelihood formats, the percentage with a range format was explored as well.

Method

Participants

The experimental sample consisted of graduate and law students attending the University of Oregon.

Procedure and Materials

Study participants were paid \$14 for approximately 1 hour of participation time. Participants were presented with simulated intelligence forecasts warning of a possible terrorist attack. Four separate terrorism scenarios were generated. The scenarios were very similar in terms of written length and the number and types of evidence used. Participants then responded to a series of questions about each scenario, filled out the numeracy individual difference measure and provided demographic information. All study procedures were passed through the University of Oregon Institutional Review Board (IRB).

Experimental Design

Unlike Study 2, this study was designed to test the specific hypotheses of interest with sufficient statistical power and in a manner that is more representative of the real environment in which consumers may view intelligence forecasts. Real consumers will most likely be looking at multiple intelligence forecasts in close proximity or directly comparing them. Therefore, a potentially better way to present the intelligence forecasts to consumers is in a within subject design.

In Study 2, consumers were sensitive to the explicit likelihood information stated in the forecasts, although consumer numeracy and the format of the likelihood information moderated these effects. However, explicit likelihood was only presented at two levels (5% and 20%) in Study 2. To further test the sensitivity of consumers to explicit statements of likelihood in intelligence forecasts, and to allow more detailed analysis of the function relating explicit to perceived likelihood, 3 levels of likelihood (i.e. 1%, 5%, 10%) were presented to consumers in Study 3. Since many forecasts in the intelligence

domain are likely to involve relatively low probability events, a lower probability range was used in this study.

In addition, explicit likelihood was not presented in a frequency format in Study 3. Understanding likelihood as a relative frequency is very natural when one is presented with an event that is repeated, like the spinning of a roulette wheel. For example, stating the probability that a particular mental patient will commit an act of violence in the next 6 months as 1/100 is clear as long as we can visualize the set to which this patient belongs: out of 100 mental patients with identical symptoms only 1 will commit an act of violence in the next 6 months. However, for many situations in which a probability value is assigned to a single event, a frequency representation of probability and therefore a frequency format is not clearly applicable. The probability of a particular act of terrorism becoming reality is a good example (see discussion in Chapter II). In addition, results from the Study 2 indicate that in the presence of a narrative description of the evidence, the frequency format did not elicit consistent ratings of risk (i.e. higher risk ratings for higher stated likelihood).

This experiment was run as a 3 (probability format/framing) x 4 (probability level) mixed experimental design with probability level as the within subject factor. The four levels of stated probability were narrative-only (no probability), 1%, 5%, 10%. The probability format factor varied as follows: 1) point estimate of probability framed as an external estimate (The probability that this event will occur is ...), 2) point estimate of probability framed as a rating of how confident the analysts are that the event is going to occur (We are x% sure that this event will occur ...) and 3) a point estimate of the probability framed as an external estimate with a confidence range around the estimate (Our best estimate of the probability that this event will occur is x%, but the probability could be as low as x% or as high as x%). The third condition is a bit different from the other two in that it includes two pieces of information — namely, the estimated external probability of the event, and an interval that gives the consumer information about how confident the analyst is in that estimate. The wider the confidence interval the less confident the analyst is in their best estimate, the narrower the interval the more

confident. Also, the pairing of scenario to probability level as well as the order of presentation were randomized to control for incidental effects.

Unlike Study 1, all of the intelligence reports were presented with a narrative description of the evidence supporting the forecast. For the evidence-only report, there was no stated probability information and there was no mention of the potential harm that could result if the attack were to occur. For the narrative reports in the probability conditions, there was a statement about the numerical probability of the event (in different formats depending on condition) as well as a statement about the potential threat or harm that would result if the attack were to occur. The statement about potential harm was held constant for all of the reports (except the narrative-only, which had no harm information): "If the attack occurs, a plausible worst-case scenario would be 1000 deaths and injuries and 50 million dollars in property damage." Since explicit likelihood and potential harm were not reported in the narrative-only forecasts, any differences between the narrative-only and numerical forecasting conditions must be interpreted as resulting from the addition of both likelihood and potential harm information.

Dependent variables

The dependent variables were designed to capture the perceived likelihood and potential harm of the forecasted terrorist plot, as well as perceptions of the usefulness and source credibility of the intelligence forecasts. As in Study 2, the first question asked about the consumer's global perceptions of risk: "How would you rate the risk associated with this possible attack?". In addition to this global question, separate questions were asked about the perceived likelihood and impressions of the overall harm or threat associated with the attack: "What is your impression of the chance that this attack will occur over the next 6 months?", "Focus on the potential outcome of the described terrorist attack. If this attack did occur, what is your impression of the overall harm that would be inflicted on people, property, the economy, etc?" In addition, consumers were asked about the perceived value or usefulness of the forecast for decision making. Finally, source credibility was assessed with a scale used by McComas & Trumbo

(2001). This measure was used to assess how consumers generally felt about the source of the intelligence forecast. The source credibility scale is made up of five questions asking the extent to which consumers trust the conclusions of the forecast, whether they feel the forecast is accurate, whether it is fair, whether it tells the whole story, and whether it is biased. Each consumer responded to the dependent variables discussed above for each of the four intelligence forecasts. Consumers responded to these dependent variables directly after reading each intelligence forecast.

After reading each of the four intelligence reports and responding to the dependent variables, the consumers were asked to make two additional ratings concerning the evidence described in each scenario. The narrative summary of the evidence had a strong influence on perceptions of the intelligence forecast presented in Study 2. In this study, consumers will be asked about specific aspects of the scenario information presented in each intelligence forecast. The first was a global rating of the overall credibility of the evidence, and the second was a rating of the how well the evidence could be formed into a coherent story. Again, these ratings were made for each of the four scenarios after the participants had finished the primary dependent variables. Numeracy was explored as an additional covariate that is stable across subjects, and the additional ratings of credibility and coherence were explored as time-dependent covariates (or within subject covariates). Participants responded to all questions on 11-point rating scales. See the Appendix C for Dependent variables.

Results

Sample Characteristics

There was a total of $n=87$ participants, resulting in 29 subjects in each between subject condition. Participants all had 4-year college degrees and the majority were current graduate/law students attending the University of Oregon. These advanced students were from a variety of departments, including biology, business, chemistry, computer science, economics, education, engineering, geological science, international

studies, law, mathematics, philosophy, physics, political science, and psychology. Tables 8 and 9 show the sample characteristics.

Table 8. Sample Characteristics.

| Characteristic | n | Mean (Median) | SD |
|------------------------------|----|------------------|------|
| Age | 87 | 28.05 (27.00) | 6.32 |
| Numeracy ^a (0-15) | 87 | 12.29 (13.00) | 2.11 |

^a Distribution is moderately negatively skewed.

Table 9. Sample Characteristics.

| Characteristic | n | % |
|---------------------------------|----|------|
| Female | 46 | 52.9 |
| Education (n=87) | | |
| 4yr college graduate | 11 | 12.6 |
| Current Graduate or Law Student | 76 | 87.4 |

The relationships between the dependent variables were examined before proceeding with the formal analysis. First, reliability analysis was conducted on the five items making up the source credibility scale (McComas & Trumbo, 2001). Inspection of scatterplots for each item pair confirmed that associations between the variables were roughly linear in nature. Reliability analyses were conducted separately for responses at each level of the within subjects variable (i.e. pure narrative, and the three numerical forecast conditions). Both the alpha coefficients ($\alpha = .850-.887$) and average inter-item correlations (average $r = .539-.592$) were sufficiently high to justify averaging the items to create a composite source credibility measure.

Table 10 shows the average Pearson correlations (averaged across the four levels of the within subjects factor) between the dependent variables related to perceptions of

chance, harm and risk in Study 3. Inspection of scatterplots for each variable pair confirmed that all of the variables were roughly linearly related.

Table 10. Average Pearson correlations w/ 95% CI's between dependent variables related to risk perception (n=87).

| | Risk | Likelihood | Harm | Credibility | Coherence |
|--------------------------|-----------------------------|------------------------------|------------------------------|-----------------------------|-----------|
| Risk | 1.00 | | | | |
| Likelihood | .587 (.710, .429) | 1.00 | | | |
| Harm | .398 (.562, .204) | .220 (.412, .010) | 1.00 | | |
| Credibility ¹ | .351 (.527, .146) | .267 (.456, .054) | .092 (.302, -.126) | 1.00 | |
| Coherence ¹ | .352 (.528, .148) | .207 (.405, -.009) | .138 (.343, -.080) | .601 (.723, .443) | 1.00 |

¹ Four cases were missing data on this variable, n=83.

Judging by the pattern of correlations in Table 10, global perceptions of risk were more strongly associated with perceptions of likelihood than perceptions of potential harm. Perceptions of the credibility and coherence of each narrative evidence summary are also significantly related to perceived risk. The focus of this study, however, is perceptions of likelihood and perceptions of global risk will not be explored further. Conceptually, there should be roughly zero correlation between perceived likelihood and harm, but a small to moderate correlation is evident in these data. Since many of the hypotheses involve expected changes in perceptions due to manipulations of stated likelihood, perceived likelihood will be the primary dependent variable.

The ratings of story coherence and evidence credibility are highly correlated. In addition, coherence and credibility show small to moderate correlations with perceptions of likelihood. Importantly, coherence and credibility correlate only weakly with potential harm, which is consistent with the theoretical model presented in the Chapter II. This makes sense because theoretically the credibility and coherence of the evidence set is pertinent to the likelihood or plausibility of the event occurring, not the potential harm.

Perceived Likelihood

The two primary goals of this study were to compare consumer perceptions of purely narrative intelligence forecasts to forecasts with explicit estimates of likelihood and potential harm, and to assess the impact of narrative and explicit likelihood information on consumer perceptions when both of these sources of information are available in a forecast. These research questions will be addressed separately below.

Pure narrative versus numerical forecasts. Two primary research questions will be explored in the analysis below:

1. Will pure narrative forecasts result in higher estimates of likelihood than forecasts with narrative and numerical estimates? Research suggests (see Chapter II) that people use scenario-based reasoning strategies that tend to inflate perceptions of likelihood when only presented with a narrative summary of the evidence relating to the target event. When numerical estimates of likelihood are presented, however, initial likelihood perceptions due to the narrative should be pulled down toward these numerical estimates.
2. Because consumers will have more difficulty evaluating likelihood in pure narrative forecasts (e.g. consumers may use idiosyncratic strategies of evaluating likelihood) than when numerical estimates are included, consumers in the pure narrative condition are expected to show more variance in perceptions of likelihood than consumers in the numerical estimate conditions. This increased variance in likelihood perceptions is an indication that the transferal of likelihood information from analyst to consumer is not as consistent in pure narrative forecasts.

Table 11 shows the effect of stated likelihood and likelihood format on consumer perceptions of likelihood. Contrary to expectations, there were no significant differences in the variance of likelihood ratings between the narrative-only and numerical conditions.

This implies that the presence of numerical estimates of likelihood and potential harm do not necessarily result in more consistent perceptions across consumers. Although there were no significant differences in the variance of likelihood ratings between the pure narrative and numerical forecasts, the distribution of the likelihood ratings were affected by the presence of the explicit numerical estimates.

Table 11. The effect of explicit numerical estimates of likelihood and potential harm and uncertainty format on consumer perceptions of likelihood.

| | Narrative | 1% | 5% | 10% | Total ^a |
|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|--------------------|
| Probability (external) | 25.60 (21.71) TB=22.99 | 19.03 (24.78) TB=5.52 | 25.17 (25.30) TB=8.20 | 23.97 (24.83) TB=9.71 | 22.72 TB=10.09 |
| Probability (internal) | 28.62 (24.71) TB=26.90 | 22.33 (24.77) TB=3.22 | 23.10 (22.66) TB=5.00 | 26.90 (23.77) TB=8.74 | 24.11 TB=19.40 |
| Probability w/range | 30.00 (22.44) TB=28.82 | 24.60 (21.55) TB=21.85 | 22.76 (20.47) TB=13.06 | 26.38 (20.13) TB=24.08 | 24.58 TB=20.09 |
| Total | 28.07 TB=24.97 | 21.99 TB=4.87 | 23.68 TB=7.43 | 25.75 TB=14.57 | |

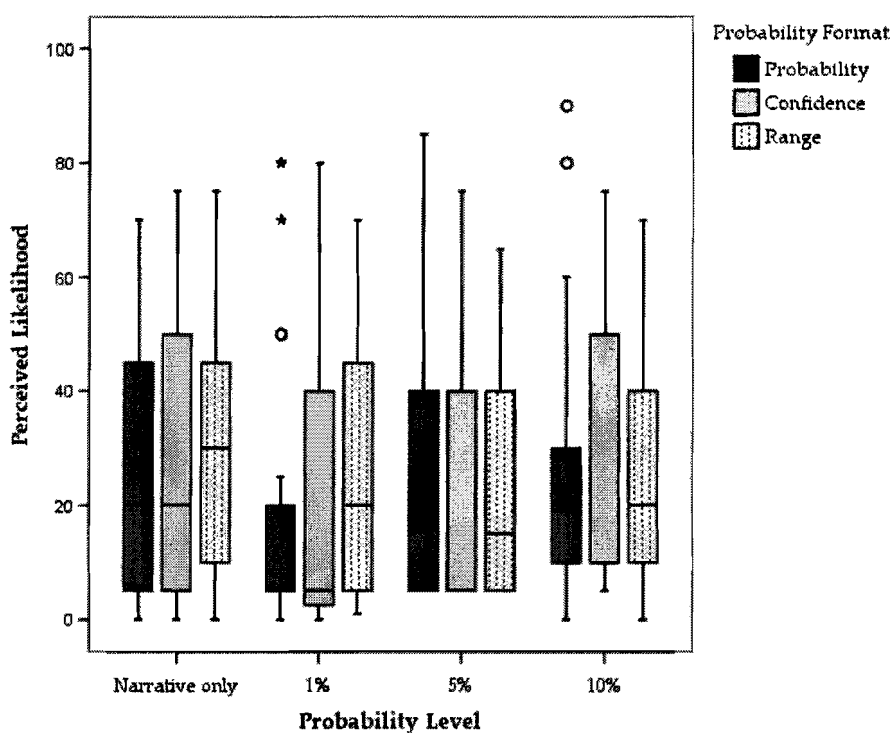
Note: There are n=29 participant ratings per cell, with N=87 total. Participants responded on a 0-100% scale. Mean (SD) and Tukey's Biweight (TB) robust measures of central tendency are reported above. Tukey's Biweight measures of central tendency provide a more robust measure of location than the mean in cases of extreme skewness and/or kurtosis (Wilcox, 2005).

^aMean totals for the between subject condition of probability format are made up of only those observations in the numerical conditions. The responses in the pure narrative condition were not included in these means because there was no explicit likelihood information present in this condition. This is necessary because the experimental design is not fully crossed.

Figure 12 shows the distributions of consumer likelihood ratings for the narrative only condition and each level of stated likelihood by likelihood format. In the narrative condition (which was identical for each level of likelihood format because no likelihood information was displayed), the distributions are slightly positively skewed, indicating that ratings tended to bunch up slightly at the low end of the probability scale. For these distributions, the mean estimates presented in Table 11 are likely to be good measures of the central tendency of the distributions. This is confirmed when comparing the mean to

the Tukey's Biweight robust measure of central tendency. Robust measures of central tendency give better estimates of central tendency in distributions with severe skewness and/or kurtosis.

Figure 12. Boxplots showing the distribution of likelihood ratings at each level of likelihood and likelihood format.



However, inspection of the boxplots for the numerical forecasting conditions reveals extreme positive skewness, particularly for the internal and external point estimate formats. For these distributions the mean is pulled sharply toward the outliers at the high end of the likelihood scale and is not a representative measure of central tendency. This is evident by comparing the means with the robust estimates of central tendency in Table 11 for these distributions. The means for these distributions are not appropriate measures of location.

This pattern of results shows that the manipulations of probability format and probability level not only changed the central tendency of the distributions but also the shape of the distributions. In addition, the shape of the pure narrative and numerical forecasting distributions are drastically different, making it very difficult to compare the means of these distributions or use any mean based statistical methods to compare the groups. For example, looking at the pattern of means in Table 11 paints a different picture than the pattern of robust measures of central tendency, the latter of which is a more accurate representation of location in the distributions. Because of the complexity introduced by the pattern of distributional changes, bootstrapping methods based on robust measures of central tendency were used to compare the pure narrative and the numerical forecasting conditions (Wilcox, 2005)⁵.

It was hypothesized that the pure narrative condition would result in higher ratings of perceived likelihood than the numerical conditions, particularly in the point estimate conditions. Inspection of the robust measures of central tendency in Table 11 shows a large difference between the narrative and numerical condition in the point estimate conditions (probability and confidence), and a smaller difference in the range conditions. In the point estimate conditions (average of probability and confidence), the narrative elicited substantially higher likelihood ratings than the 1% (Diff = 20.77, 95%CI: 1.28, 26.67), 5% (Diff = 17.32, 95%CI: -4.69, 24.14), and 10% (Diff = 16.01, 95%CI: -2.60, 21.84) numerical conditions. In the range condition the differences between the narrative and numerical conditions were much smaller, 1% (Diff = 7.01, 95%CI: -0.76, 24.64), 5% (Diff = 14.38, 95%CI: 2.62, 24.82), and 10% (Diff = 5.18, 95%CI: -4.96, 18.28)⁶.

⁵ It is difficult to transform the distributions of likelihood ratings to a more normal shape because the skewness is not consistent across the distributions. For example, a log transformation to correct the skewness of the likelihood ratings at one condition will bias distributions that are relatively normal in the opposite direction, and since all of the scales on the repeated measures factor must be consistent, different transformation cannot be applied to different distributions.

⁶ In general, bootstrap methods tend to be less powerful as compared to standard mean-based statistics, and although the differences reported are substantial in magnitude, some comparisons were not significant at $\alpha=.05$ (although all $p's < .10$). These same comparisons were also conducted with standard mean-based methods, and although the mean is not an accurate measure of location in these distributions (discussed above), all group comparisons were significant at $\alpha=.05$.

Overall, pure narrative forecasts resulted in greater consumer perceptions of likelihood than forecasts with 1%, 5%, and 10% numerical likelihood estimates. This suggests that consumers were sensitive to the explicit likelihood information in the forecasts and they, consequently, lowered their ratings of likelihood from what would be estimated from the narrative information alone. In the next section, precisely how consumers used both the explicit likelihood and narrative information is explored in more detail.

The impact of explicit likelihood and narrative information on perceptions of likelihood. Several research questions are explored in this section:

1. In the presence of a narrative evidence summary, to what extent will consumers use the explicit likelihood information to inform their perceptions of likelihood? Forecasts that include higher explicit estimates of likelihood should result in higher perceived likelihood on the part of consumers.
2. In the presence of explicit numerical estimates of likelihood, to what extent will consumers use the coherence and credibility of the evidence in the narrative summary to inform their perceptions of likelihood? Consumers that perceive greater coherence and credibility in the narrative summary will perceive greater likelihood.
3. To what extent will the format of the likelihood information moderate the effect of explicit likelihood and narrative information on perceptions of likelihood?
 - Does presenting the likelihood as an internal confidence rating as compared to an external likelihood affect consumer's sensitivity to this information?
 - Consumers may show less sensitivity to the explicit likelihood information when likelihood is presenting with a range of plausible value. The results from Study 2 suggest that this may be the case.

4. To what extent will consumer numeracy moderate the effect of explicit likelihood and narrative information on perceptions of likelihood?
 - Consumers lower in numeracy may not be as sensitive to numerical likelihood information and may focus more on the narrative summary when making likelihood judgments. The opposite may be true for consumers higher in numeracy.

Explicit likelihood: One of the primary research questions in this study was how consumers used the explicit likelihood information to inform their perceptions of likelihood. For example, if consumers completely ignored the scenario information and directly translated the explicit likelihood estimates to their rated perceptions of likelihood (i.e. 1% stated, 1% reported; 5% stated, 5% reported, etc), the slope relating stated to perceived likelihood should be roughly equal to 1 (assuming that consumers were using the best estimate in the point estimate w/range format). This assumes, however, that participants were using the rating scale as a percentage likelihood scale, and they were not using it as a more generic scale in which they tried to scale their feelings of likelihood in terms of relative magnitude⁷. Examination of the distributions of the likelihood ratings in Figure 5.11 shows that consumers were clearly not directly translating stated likelihood into rated perceived likelihood. The linear function relating stated likelihood to rated perceived likelihood had an intercept = 21.58 and a slope = 0.417 (significantly different from $b = 1$, $p < .05$).

It is clear that consumers were not directly translating stated likelihood into perceived likelihood. This either indicates that these consumers were interpreting the probability scale appropriately (0-100% chance) and they were simply using other information from the forecast to make their likelihood ratings, or they were not using the probability scale in the strict sense and simply using it as a generic rating scale, in which they tried to scale

⁷ In addition, the likelihood rating scale was presented in 5-point steps (i.e. 0-5-10-15 etc), and although many consumers reported values between the steps (e.g. 1%), the crudeness of the scale may have affected how consumers reported perceptions of likelihood.

their feelings of likelihood in terms of relative magnitude. However, even if consumers were using the rating scale as a generic scale in which they tried to scale their feelings of likelihood, they should still make ordinal differentiations between the levels of stated likelihood. One would expect a 10% stated likelihood to be rated higher than a 5% stated likelihood which would be rated higher than a 1% likelihood. Thus, even if consumers were not all using the likelihood rating scale in the same way, the extent to which consumers adhere to a monotonic relationship between stated and perceived likelihood will be an indication of their sensitivity to the stated likelihood information.

Only 39.08% of consumers showed a consistent monotonic relationship among their rated perceptions of likelihood as stated likelihood increased⁸. Thus, approximately 60% of the consumers in this sample were not consistently perceiving higher likelihood as the stated likelihood increased. Perhaps the stated likelihood values of 1% and 5% were too small and close together to be distinguished by many consumers, and the requirement for sensitivity to the explicit likelihood should be loosened even more – namely, that only a stated likelihood of 10% must result in higher perceptions of likelihood than a stated likelihood of 1%. A much larger percentage of consumers, 67.82%, perceived greater likelihood in forecasts with a 10% stated likelihood as compared to a 1% stated likelihood. Although the consumers were, to some extent, using stated likelihood to inform their perceptions of likelihood, they appeared to be using other information as well. In the next section, the extent to which consumers also use the properties of the narrative evidence summary to inform their perceptions of likelihood is explored. Also of interest are the potential moderating effects of the format of the likelihood information and consumer numeracy level.

⁸ As discussed above, the likelihood rating scale was presented to consumers in 5-point intervals (i.e. 0%-5-10-15, etc). Although some consumers reported values in between these 5-point intervals (e.g. 1%), this crude measurement scale may have affected ratings of perceived likelihood. For example, a consumer may have tried to directly translate stated likelihood (e.g. 1%, 5%, 10%) to rated perceived likelihood, but because there was not an explicit 1% on the rating scale, he or she may have reported 5%, 5%, 10%. Thus, if a consumer reported this pattern of likelihood ratings they were given credit for having made an ordinal differentiation (i.e. monotonic function) among the levels of stated likelihood.

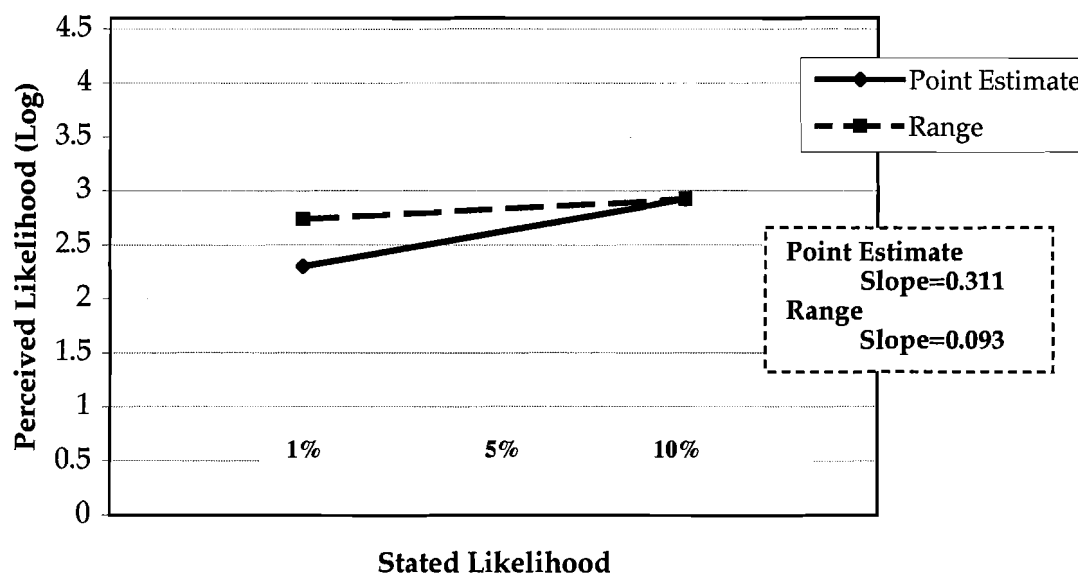
Explicit likelihood and narrative information: The next set of hypotheses were tested in the context of several multilevel mixed models, with the repeated measures represented at level 1 (i.e. within subject manipulations and covariates varying within subjects), and the between-subject data represented at level 2 (i.e. between subject manipulations, and subject-level covariates). See Appendix E for statistical details concerning the multilevel models used to estimate effects in Study 3 and Study 4.

The primary goals of this analysis are to test the hypothesized effects of stated likelihood information and properties of the narrative summary on perceptions of likelihood. In addition, the potentially moderating influence of likelihood format and consumer numeracy will be explored. These hypothesized relationships were discussed in Chapter III (see Model of Consumer Perceptions of Intelligence Forecasts). The direct effect of stated likelihood on perceived likelihood has already been discussed above, although in these analyses likelihood format and numeracy will be explored as moderators of this effect. In these analyses, the effect of stated likelihood is represented as the linear slope across the stated probability levels (higher order polynomial effects were not significant), which indexes the extent to which consumers perceived greater likelihood as stated likelihood increased. As noted above, credibility and coherence ratings were highly correlated with one another, and including them in the multilevel models resulted in moderate multicollinearity problems. Several models were fit with coherence and credibility modeled separately, and the results were comparable. On the basis of the similar pattern of relationships between the variables and model parsimony, the credibility and coherence ratings were averaged to create a composite variable that will be called “evidence properties”. This composite variable can be conceptualized as the extent to which each consumer found the evidence in each scenario to be credible and coherent. In addition, the likelihood ratings were log transformed in all analyses to reduce the skewness problems discussed above.

In the first model, ratings of perceived likelihood were modeled as a function of stated likelihood and likelihood format. The linear function relating stated likelihood to perceived likelihood was positive, as detailed above, and significantly different from 0,

slope = 0.239, $t(255) = 3.996$, $p < .001$, $ES = .53$ ⁹. The primary goal of this analysis, however, was to test if the format of the likelihood information moderated this effect. The likelihood format did reliably explain some of the variance in the slopes relating stated to perceived likelihood, $t(255) = -1.859$, $p = .06$, $ES = .16$, such that consumers in the range condition showed flatter slopes than consumers in the point estimate conditions (i.e. internal and external point estimates). Consumers in the range condition showed less differentiation between the levels of stated likelihood. There were no significant differences between the internal and external point estimate conditions. Figure 13 shows the effect of stated likelihood on perceived likelihood for the point estimate and range conditions.

Figure 13. The effect of stated likelihood on perceived likelihood for the point estimate and range conditions.



This result demonstrates that stated likelihood had a larger linear effect on perceived likelihood in the point estimate conditions than in the range condition. Overall, however, there was still quite a bit of variability in perceived likelihood that was not accounted for

⁹ See Appendix E for discussion of the effect size metric used for the HLM results.

by stated likelihood or likelihood format. In the next set of models the coherence and credibility of the narrative evidence and individual differences in numeracy were added as further predictors of perceived likelihood.

In the full model, perceptions of likelihood were modeled as a function of stated probability and perceptions of credibility and coherence of the narrative summary at the first level. At the second level, contrasts comparing the likelihood format conditions and consumer numeracy were added. Table 12 shows the results for the full multilevel model. Each effect will be described in more detail below.

The first hypothesis tested was that consumer perceptions of the credibility and coherence of the narrative evidence summary would relate to perceptions of the likelihood of the target event. Higher ratings of coherence and credibility were found to relate to higher perceived likelihood, $t(237) = 4.740$, $p < .001$, $ES = 1.06$. As above, there was also a significant linear effect of stated likelihood on perceived likelihood, $t(237) = 3.245$, $p = .002$, $ES = 0.54$. These results suggest that the perceived properties of the narrative evidence had a larger effect on perceived likelihood than the manipulations of stated likelihood (i.e. 1%, 5%, 10%).

Table 12. Multilevel model results for perceived likelihood.

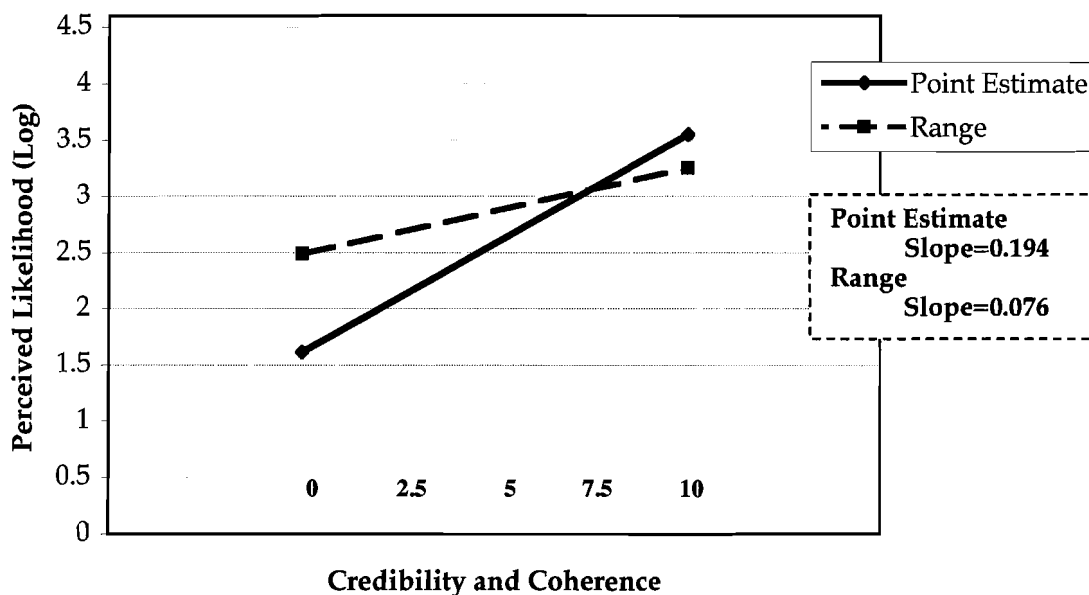
| Fixed Effect | Coefficient | SE | t | p-value |
|------------------------------------|-------------|------|--------|-----------------|
| Mean likelihood (Intercept) | | | | |
| Mean intercept (Intercept) | 2.64 | 0.09 | 29.10 | <.001 |
| Range vs Point estimates | 0.10 | 0.06 | 1.583 | .117 |
| Prob vs Confidence | 0.05 | 0.11 | 0.430 | .668 |
| Numeracy | -0.15 | 0.05 | -2.90 | .005 |
| Stated Likelihood (Slope) | | | | |
| Mean slope (Intercept) | 0.17 | 0.05 | 3.245 | .002 |
| Range vs Point estimates | -0.04 | 0.03 | -1.383 | .168 |
| Prob vs Confidence | -0.04 | 0.07 | -0.601 | .548 |
| Numeracy | 0.10 | 0.02 | 4.162 | <.001 |
| Evidence Properties (Slope) | | | | |
| Mean slope (Intercept) | 0.15 | 0.03 | 4.740 | <.001 |
| Range vs Point estimates | -0.04 | 0.02 | -2.237 | .026 |
| Prob vs Confidence | 0.04 | 0.04 | 1.030 | .305 |
| Numeracy | -0.03 | 0.01 | -2.095 | .037 |

Note: The three levels of likelihood format were tested as two orthogonal helmert contrasts. Contrast 1 compared the range condition to the average of the two point estimate conditions. Contrast 2 compared the two point estimate conditions to each other (probability as an external estimate versus probability as an internal, subjective confidence statement).

In addition, the format of the likelihood information significantly moderated the effect of evidence properties on perceived likelihood. Consumers in the point estimate conditions used the evidence properties to rate perceived likelihood more than the consumers in the range condition, $t(237) = -2.237$, $p = .037$, $ES = 0.23$. This is counter to expectation, in that one might expect the consumers to use the evidence properties more in the range condition because there was no single probability from which to judge likelihood. In fact, previous research has suggested that when presented with a

confidence range, participants are more likely to ignore that information and focus on other information to make the judgment at hand. Figure 14 shows the effect of perceptions of coherence and credibility on perceived likelihood for consumers in the point estimate and range conditions.

Figure 14. The relationship between perceived credibility/coherence and perceived likelihood for consumers in the point estimate and range conditions.



The Effect of Numeracy: Participants higher in numeracy reported greater levels of perceived likelihood, $t(79) = -2.902$, $p = .005$, $ES = .37$. In Study 2, a similar relationship was observed between numeracy and perceived risk. Numeracy also moderated the effect of the stated likelihood and the evidence properties on perceived likelihood. The effect of stated likelihood on perceived likelihood was smaller for consumers lower in numerical ability, $t(237) = 4.162$, $p < .001$, $ES = .56$, and the relationship between the ratings of the evidence and perceived likelihood was higher for consumers lower in numeracy, $t(237) = -2.095$, $p = .037$, $ES = .22$. These results suggest that consumers lower in numeracy were more sensitive to the perceived properties of the narrative

information and less sensitive to the stated probability information, which follows if the stated probability information was not as evaluable and more difficult to use for those lower in numerical ability. Figures 15 and 16 show the effects of stated likelihood and evidence properties on perceived likelihood for consumers with different levels of numeracy.

Figure 15. The effect of stated likelihood on perceived likelihood for consumers with different levels of numeracy.

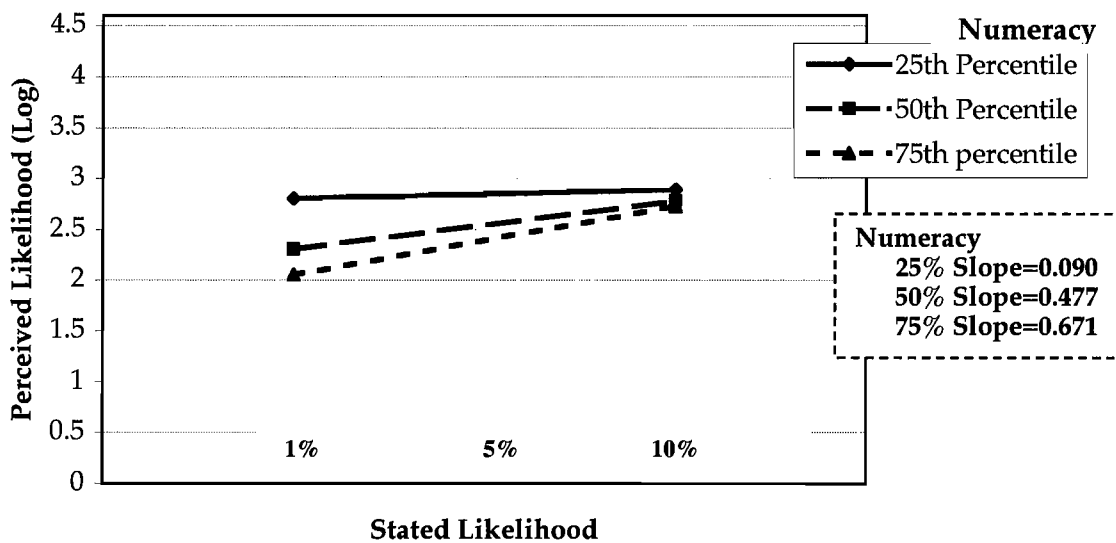
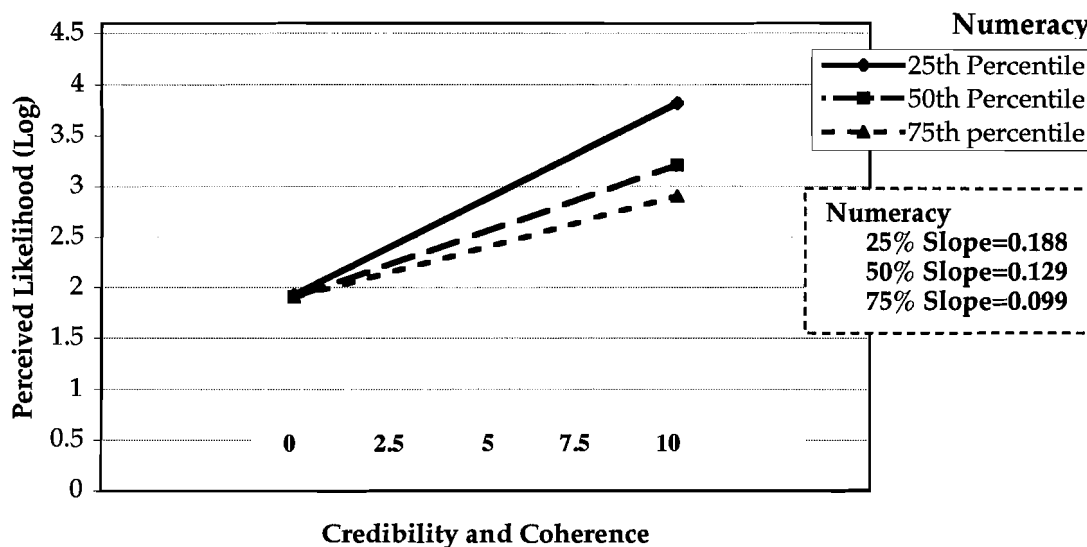


Figure 16. The relationship between perceived credibility/coherence and perceived likelihood for consumers with different levels of numeracy.



Perceived usefulness and source credibility

The primary research questions concerning perceived usefulness and source credibility are presented below:

1. Will consumers feel that a forecast is more useful and has higher source credibility when explicit numerical estimates of likelihood and potential harm are presented?
2. Will consumers feel that a forecast is more useful and has higher source credibility when they perceive there to be greater coherence and credibility in the narrative evidence summary?
3. Will the format of the likelihood information affect perceived usefulness and source credibility?
4. Will numeracy affect the perceived usefulness and source credibility of intelligence forecasts?

The source credibility composite measure and perceptions of usefulness showed a moderate to strong correlation, $r = .534$, 95% CI (.669, .364). Perceived usefulness and source credibility are modeled separately below¹⁰.

Perceived Usefulness. Ratings of perceived usefulness, or how valuable the intelligence forecasts were for decision making, were modeled as a function of stated likelihood, likelihood format, perceptions of the credibility and coherence of the narrative and numeracy. As expected, the pure narrative condition elicited lower perceived usefulness than the numerical conditions, $t(319) = -1.978$, $p = .048$, $ES = .23$, and the internal (confidence) likelihood format tended to elicit lower levels of usefulness than the probability and range conditions combined, $t(79) = -1.735$, $p = .086$, $ES = .17$. Additionally, as hypothesized, consumers who rated the evidence set as more coherent and credible reported higher levels of perceived usefulness, $t(319) = 11.950$, $p < .001$, $ES = 2.78$. Finally, consumers higher in numeracy reported lower levels of perceived usefulness than consumers lower in numeracy, $t(79) = -2.623$, $p = .011$, $ES = .30$.

¹⁰ Since these two variables are correlated, a multivariate analytic framework would be ideal for this analysis. Since multivariate multilevel models are still relatively new and difficult to estimate at this point, other analyses were conducted to assess the independent effects of these variables (i.e. removing the shared variance). For example, a separate multilevel model was estimated for the residualized source credibility measure (i.e. the variance in the source credibility measure that could not be explained by perceived value, acquired by regressing source credibility on perceived value and saving the residuals). There were no substantive differences between the results with the residualized variables and the results with the full variables reported below.

Table 13. The effect of stated likelihood and likelihood format on perceived usefulness.

| | Narrative | 1% | 5% | 10% | Total ^a |
|---------------------------|---------------------|---------------------|---------------------|---------------------|--------------------|
| Probability (external) | 4.36 (2.60) n=28 | 4.64 (2.54) n=28 | 4.93 (2.51) n=28 | 4.57 (2.28) n=28 | 4.71 n=28 |
| Probability (internal) | 3.55 (2.06) n=29 | 3.83 (2.29) n=29 | 4.45 (2.05) n=29 | 4.45 (2.18) n=29 | 4.24 n=29 |
| Probability w/range | 4.93 (1.98) n=28 | 4.93 (1.72) n=28 | 4.61 (1.64) n=28 | 4.82 (1.63) n=28 | 4.78 n=28 |
| Total | 4.27 n=85 | 4.46 n=85 | 4.66 n=85 | 4.61 n=85 | |

Note: Mean (SD) and sample size (n) are reported.

^a Mean totals for the between subject condition of probability format are made up of only those observations in the numerical conditions. In other words, the responses in the pure narrative condition were not included in these means because there was no explicit probability information present in this condition. This is necessary because the experimental design is not fully crossed.

Source Credibility. Perceived source credibility was modeled as a function of stated likelihood, likelihood format, perceptions of the credibility and coherence of the narrative and numeracy. The first hypothesis tested was that consumers would perceive greater source credibility in the numerical forecasting conditions as opposed to the pure narrative condition. There was virtually no difference between these conditions (see Table 14). The next hypothesis was confirmed, that consumers who perceived greater coherence and credibility in the evidence set would also perceive greater overall source credibility, $t(322) = 9.847, p < .001, ES = 1.84$. Individual differences in numeracy were not significantly related to perceptions of source credibility.

Table 14. The effect of stated likelihood and likelihood format on perceived source credibility.

| | Narrative | 1% | 5% | 10% | Total ^a |
|------------------------|---------------------|---------------------|---------------------|---------------------|--------------------|
| Probability (external) | 5.23 (1.74) n=28 | 5.18 (1.96) n=28 | 5.60 (1.80) n=28 | 5.11 (2.00) n=28 | 5.30 n=28 |
| Probability (internal) | 4.78 (2.06) n=29 | 5.20 (2.02) n=29 | 4.92 (1.97) n=29 | 5.14 (2.14) n=29 | 5.09 n=29 |
| Probability w/range | 5.29 (1.59) n=28 | 5.08 (1.50) n=28 | 4.77 (1.48) n=28 | 4.98 (1.48) n=28 | 4.94 n=28 |
| Total | 5.09 n=85 | 5.16 n=85 | 5.10 n=85 | 5.08 n=85 | |

^a Mean totals for the between subject condition of probability format are make up of only those observations in the numerical conditions. In other words, the responses in the pure narrative condition were not included in these means because there was no explicit probability information present in this condition. This is necessary because the experimental design is not fully crossed.

Summary and Discussion

Overall, the results from this experiment reveal important differences in likelihood perception and perceptions of usefulness and source credibility between narrative-only forecasts and those with explicit probability and potential harm information added to the narrative forecast. In addition, when presented with both narrative and numerical information with which to judge likelihood and harm, consumers appeared to be more greatly affected by their perceptions of the credibility and coherence of the narrative than the explicit likelihood information (narrative, $ES = 1.06$; stated likelihood, $ES = 0.54$), and these effects were moderated by the format of the probability information and the numerical ability of the consumer.

Pure narrative forecasts versus numerical forecasts

There are two potential communication problems when communicating risk information from analyst to consumer. The first is that different consumers may not perceive the same levels of likelihood or risk in a forecast (i.e. there will be large amount

of variance in judgment), and the second is that consumers may be systematically biased in their judgments (e.g. consistently perceiving more likelihood or risk than the analyst intended). It was hypothesized that the presence of explicit estimates of likelihood would facilitate more consistent transfer of likelihood information from analyst to consumer and reduce the idiosyncratic ways in which likelihood is estimated from a narrative evidence summary. Contrary to expectation, there were no differences in the variance of likelihood ratings between the pure narrative and numerical forecasting conditions, although there were other effects on central tendency and the shape of the distributions of likelihood ratings. The narrative-only condition resulted in higher estimates of likelihood than the numerical forecasting conditions. Presumably, when given only narrative information with which to judge risk or likelihood, estimates tend to be inflated due to the scenario-based reasoning processes that pure narrative forecasts elicit. For example, consumers may judge the likelihood of the target event by the plausibility of the terrorist scenario, using representative and simulation type heuristics that overwhelm statistical thinking about the problem (discussed in Chapter II). When given the explicit numerical likelihood and harm information, however, these figures act as an anchor or frame with which to judge the likelihood of the scenario. Thus, consumers initial perceptions of likelihood based on the narrative summary of the evidence were pulled down toward the explicit likelihood estimates presented in the forecast. In addition, and as expected, forecasts with only a narrative summary were judged as less useful for decision making than forecasts with numerical estimates of likelihood and potential harm.

The relationship between stated likelihood and perceived likelihood

We have seen that consumers were sensitive to the explicit likelihood information in the forecasts, but how did they use this information to inform their perceptions of likelihood? It is clear from examining the distributions of perceived likelihood that the majority of consumers were not directly transferring stated likelihood into perceived likelihood. One would expect, however, that the explicit likelihood estimates presented by the analyst were at least differentiated in an ordinal fashion in consumer's perceptions

of likelihood. However, only 39.08% of consumers showed a monotonic relationship between stated and perceived likelihood (i.e. perceptions likelihood for stated 10% > 5% > 1%), and 67.82% of consumers showed greater perceptions of likelihood for forecasts with 10% stated likelihood as opposed to 1% stated likelihood. In addition, the general trend for consumers to perceive greater likelihood as the stated likelihood increased was moderated by the format of the likelihood information. This effect was stronger in the point estimate conditions than in the range condition. Presumably, in the point estimate condition there is a single number that consumers could use to inform perceptions of likelihood, while in the range condition this was not the case. In summary, consumers were sensitive to the explicit likelihood information presented in the forecasts, although many consumers appeared to be using other information to inform their perceptions of likelihood as well.

The effect of both stated probability and properties of the narrative summary

Even in the presence of explicit likelihood estimates, consumers that perceived greater coherence and credibility in the narrative summary reported greater perceptions of likelihood. In fact, perceptions of credibility and coherence had a larger effect on perceived likelihood than the manipulation of stated likelihood (1%, 5%, 10%). In addition, higher perceived coherence and credibility also related to higher perceived usefulness and source credibility of the forecast.

This result fits in well with previous research showing that the interpretation of numerical expressions of likelihood are affected by contextual information (Windschitl and colleagues, see Chapter II). Even though one might expect numerical expressions of likelihood to be unambiguously interpreted because they are precise, many consumers have trouble evaluating these estimates to inform their perceptions of likelihood. In this case, consumers also used the narrative evidence summary to inform their perceptions of likelihood. In addition, one might expect consumers who are better able to evaluate numerical estimates of likelihood to use the stated likelihood estimates and be less affected by the narrative evidence summary. Conversely, consumers who have difficulty

evaluating numbers will not be able to use the stated likelihood to inform their perceptions of likelihood and will focus more on other contextual information like the narrative evidence summary. The results showed precisely this effect. Consumers lower in numeracy tended to use the narrative evidence summary to judge likelihood and showed less sensitivity to the stated likelihood information. Consumers higher in numeracy showed the opposite pattern.

Other effects of likelihood format

Consumers presented with the probability point estimates showed a larger relationship between ratings of the narrative evidence and likelihood judgments than those presented with the range condition. This result was contrary to expectation. Previous research suggests that when participants are presented with a range of values, they will tend to ignore the information and use other more easily evaluated information to make the judgment at hand (Hsee, 1995; see Chapter II). In this case, one might expect that consumers would be more likely to use the narrative information in the range condition as compared to the point estimates conditions. In fact, it appears that consumers were less influenced by their perceptions of the credibility and coherence of the evidence in the range condition.

In addition, when probabilities were expressed as a confidence rating (i.e. "...we are x% sure that this attack will occur over the next six months), consumers found them to be less valuable than when they were presented as external probabilities (i.e. "...the probability of this attack occurring over the next six months is x%") or as external probabilities with a range.

Study 4 – Exploring Consumer Perceptions of Intelligence Forecasts in Hindsight.

Purpose

The primary focus of Study 4 is to explore how consumers feel about the source credibility and usefulness of intelligence forecasts in hindsight, and to what extent consumers assign blame to forecasters after knowing the outcome of a forecasted event.

Of particular interest are the types of information that consumers use to make source credibility, usefulness, and blame judgments in hindsight. For example, when evaluating a forecast in hindsight, will consumers still be sensitive to the stated likelihood information and credibility and coherence of the narrative evidence summary?

One of the factors that might make the intelligence community reluctant to attach numerical probability estimates to their analytic judgments is that they feel they could be more readily blamed if a forecasted low likelihood event occurs or a high likelihood event does not occur. For example, consumers may blame the analyst for providing a poor forecast if an event assigned a likelihood of 5% occurs. The analysts may feel more insulated from blame if they keep things vague and non-falsifiable, which could be one reason why current intelligence reporting is primarily narrative in nature (Schrage, 2005).

When forecasting potential terrorist threats, it seems that an analyst would be more afraid of the perceived “error” in which they make a forecast with a low probability and then the attack occurs. The opposite “error”, a high probability forecast of an attack that does not occur, would likely receive less attention precisely because the feared event did not occur. In this study, the focus is on consumer perceptions of intelligence reports in which an analyst assigns a relatively low probability to an event that eventually does occur within the specified timeframe of the forecast.

Method

Participants

Participants were a mix of undergraduate students from the University of Oregon and members of the community with undergraduate college degrees.

Procedure and Materials

Study participants were paid \$14 for approximately 1 hour of participation time. Participants were presented with the same simulated intelligence reports used in Study 2 (see Appendix D for materials). In this study, however, consumers first read a brief

passage about a terrorist attack that occurred a few weeks ago. A passage from one of the scenarios is presented below:

“Summary of the attack:

Several weeks ago, a bomb was detonated on a passenger ship in New York City killing over 900 people and wounding hundreds more. It has become clear that militant group ZZZ was responsible for the attack. The attack would most likely have been stopped if additional security had been assigned to protect targets in New York City. A special congressional committee has been formed and several politicians have begun criticizing the intelligence community.

Turn to the next page to read an intelligence report that was submitted to senior decision makers three weeks before this attack. You will then be asked to make a series of judgments about this intelligence report.”

Participants then read an intelligence report that was written a few weeks before the attack occurred. They then made a series of judgments about the report. All study procedures were passed through the University of Oregon Institutional Review Board (IRB).

Experimental Design

This experiment was run as a 3 (probability format/framing) x 4 (probability level) mixed experimental design with probability level as the within subject factor. The experimental design was identical to the design used in Study 3.

Dependent variables

For the most part, the dependent variables were identical to Study 3, although in this study they are framed in hindsight (see Appendix D). However, there was one additional question that asked participants to rate the amount of blame that they felt the forecasters deserved – “Think about both the intelligence report and the terrorist attack that occurred

three weeks later. Some people are blaming the intelligence community for not doing a good job predicting whether this attack would occur. How much blame do you think should be placed on the analysts that produced the intelligence report?" Participants responded to all questions on 11-point rating scales.

Results

Sample Characteristics

There was a total of n=81 participants, resulting in 27 subjects in each between subject condition. The majority of the participants were undergraduate students, although approximately 25% were either current graduate students or had 4-year college degrees. Tables 15 and 16 show the sample characteristics.

Table 15. Sample Characteristics

| Characteristic | n | Mean (Median) | SD |
|------------------------------|----|------------------|------|
| Age ^a | 80 | 23.43 (21.00) | 7.73 |
| Numeracy ^b (0-15) | 81 | 11.43 (12.00) | 2.20 |

^a One participant did not report age.

^b Distribution is moderately negatively skewed.

Table 16. Sample Characteristics

| Characteristic | n | % |
|-------------------------------|----|------|
| Female ^a | 40 | 49.4 |
| Education ^a (n=80) | | |
| Some College | 60 | 74.1 |
| 4yr college graduate | 18 | 22.2 |
| Current Graduate/Law Student | 2 | 2.5 |

^a There was one case that did not indicate sex or education level.

Before proceeding with the formal analysis, the relationships between the dependent variables were examined. First, reliability analysis was conducted on the five items making up the source credibility scale (McComas, 2001). Reliability analyses were conducted separately for responses at each level of the within subjects variable (i.e. pure narrative, and the three numerical forecast conditions), and as in Study 3, both the alpha coefficients ($\alpha = .839-.903$) and average inter-item correlations (average $r = .523-.668$) were sufficiently high to justify averaging the items to create a composite source credibility measure.

Table 17 shows the average correlations (averaged across the four levels of the within subjects factor) between perceived blame, perceived value, and source credibility. Table 18 shows the average Pearson correlations between the dependent variables related to perceptions of likelihood, potential harm and risk. Inspection of scatterplots for each variable pair confirmed that all of the variables were roughly linearly related.

Table 17. Average Pearson correlations w/ 95% CI's between blame, usefulness, and source credibility (n=81).

| | Blame | Usefulness | Source Cred |
|-------------|-------------------------------|-----------------------------|-------------|
| Blame | 1.00 | | |
| Usefulness | .092 (.304, -.129) | 1.00 | |
| Source Cred | -.157 (.064, -.363) | .586 (.713, .422) | 1.00 |

Table 18. Average Pearson correlations w/ 95% CI's between dependent variables related to risk perception (n=87).

| | Risk | Chance | Harm | Credibility | Coherence |
|--------------------------|-----------------------------|-----------------------------|------------------------------|-----------------------------|-----------|
| Risk | 1.00 | | | | |
| Chance | .683 (.784, .546) | 1.00 | | | |
| Harm | .487 (.638, .301) | .360 (.536, .154) | 1.00 | | |
| Credibility ¹ | .519 (.664, .337) | .420 (.585, .222) | .220 (.418, .002) | 1.00 | |
| Coherence ¹ | .462 (.619, .270) | .357 (.534, .150) | .205 (.405, -.014) | .733 (.820, .613) | 1.00 |

¹ Two cases were missing data on this variable, n=79.

Overall, the pattern of correlations is roughly consistent with those reported in Study 3. Again, perceived likelihood was more highly correlated with global perceptions of risk than perceptions of potential harm. In addition, there was a small to moderate correlation between perceived likelihood and perceived harm.

The ratings of story coherence and evidence credibility are highly correlated, and coherence and credibility show moderate correlations with perceptions of chance, which is consistent with the theoretical model presented in Chapter III. As in Study 3, coherence and credibility show smaller correlations with perceived potential harm.

Finally, the correlation between perceived usefulness and source credibility was consistent with the pattern observed in Study 3. Perceived blame was not significantly correlated with the other perception variables.

Perceived Blame

Several research questions were explored below:

1. The primary comparison of interest is between the pure narrative and numerical estimate conditions. Will the purely narrative forecast elicit less blame than a forecast with a low probability estimate assigned to an event that eventually occurs?
2. Within the numerical forecasting conditions, consumers may assign more blame the lower the estimated probability of the attack, perceiving these forecasts to be more “wrong”.
3. The probability point estimate (which appears more precise) may elicit more blame than the point estimate with range. In addition, consistent with previous research (Fox and Malle, 1995), participants may assign more blame when the probability point estimate is framed as a confidence rating (internal) as opposed to an external probability.
4. Participants may also use their perceptions of the credibility and coherence of the narrative evidence summary to inform their perceptions of blame.

Table 19 shows the effect of stated likelihood and likelihood format on perceptions of blame. The distributions of perceived blame ratings were not skewed, and there were no missing data.

Table 19. The effect of stated likelihood and likelihood format on perceptions of blame.

| | Narrative | 1% | 5% | 10% | Total ^a |
|---------------------------|---------------------|---------------------|---------------------|---------------------|--------------------|
| Probability (external) | 5.19 (3.07) n=27 | 5.19 (2.40) n=27 | 5.11 (2.55) n=27 | 4.85 (2.43) n=27 | 5.05 n=27 |
| Probability (internal) | 4.17 (2.97) n=27 | 5.13 (2.50) n=27 | 4.80 (2.63) n=27 | 4.41 (2.31) n=27 | 4.78 n=27 |
| Probability w/range | 4.67 (3.17) n=27 | 4.69 (2.64) n=27 | 3.70 (2.33) n=27 | 4.04 (2.78) n=27 | 4.14 n=27 |
| Total | 4.67 n=81 | 5.00 n=81 | 4.54 n=81 | 4.43 n=81 | |

^a Mean totals for the between subject condition of probability format are made up of only those observations in the numerical conditions.

Effects of explicit likelihood. The analysis proceeded in a similar manner as Study 2, in which multilevel models were used to model the effect of the experimental manipulations, perceptions of the credibility and coherence of the evidence summary and numeracy. The first research question of interest was whether perceived blame differed between the pure narrative and the average of the numerical forecasting conditions. There was no significance difference between these conditions in terms of perceived blame. Within the numerical forecasting conditions, however, there was a significant linear trend across the probability levels, $t(237) = -1.974$, $p = .049$, $ES = .48$, such that consumers reported less blame as stated likelihood increased. The numerical probability information was perceived as a relevant source of information for judging blame, and they were interpreting the forecasts with the lower probabilities as more “wrong”.

Effects of likelihood format. There was also a trend for consumers to report less blame in the range condition than in the two point estimate conditions, $t(78) = -1.661$, $p = .10$, $ES = .14$. However, since consumers were sensitive to the stated likelihood information when making judgments of blame, this effect may be due to the fact that the range format provided a range that included higher estimates of likelihood (e.g. in the 10% condition it ranged to as much as 30% at the high end). However, the range

condition elicited lower perceived blame even as compared to point estimates that equaled the high end of the provided range (e.g. compare the mean blame in the point estimate conditions for 10% with the mean blame in the 5% condition for the range conditions, which provided the following interval, High: 10%, Best: 5%, Low: 1%). In other words, the observed difference between the point estimate and range conditions cannot be fully explained by the range format providing higher estimates of probability within the provided intervals. In addition, there were no significant differences in perceived blame between the internal and external framing of likelihood, $t(78) = 0.491$, $p = .625$.

Evidence properties, perceived harm, and numeracy. Contrary to expectation, ratings of the coherence and credibility of the narrative evidence did not significantly relate to perceived blame, $t(308) = -1.309$, $p = .192$. However, perceptions of potential harm were significantly related to perceptions of blame, $t(308) = 2.883$, $p = .005$, $ES = .55$, such that greater perceptions of harm were associated with greater ratings of blame. Numeracy had an overall effect on perceived blame, $t(77) = -2.399$, $p = .019$. $ES = .29$, such that consumers lower in numeracy reported that more blame should be placed on the analysts.

Perceived Usefulness and Source Credibility

The research questions explored in this section are detailed below:

1. When examining an intelligence forecast in hindsight, will participants perceive different levels of usefulness and/or source credibility between reports with only narrative versus those with numerical estimates?
2. Will consumers be sensitive to the stated likelihood and properties of the narrative summary when judging usefulness and source credibility in hindsight?
3. Will numeracy or the format of the stated likelihood affect perceptions of usefulness or source credibility?

As in Study 3, perceived usefulness and source credibility are modeled separately. The distributions of perceived value and source credibility were not grossly skewed, and there were no missing data for perceived value and one missing case on the source credibility measure. Tables 20 and 21 show the effects of stated likelihood and likelihood format on perceptions of usefulness and source credibility.

Table 20. The effect of stated likelihood and likelihood format on perceived usefulness.

| | Narrative | 1% | 5% | 10% | Total ^a |
|---------------------------|---------------------|---------------------|---------------------|---------------------|--------------------|
| Probability (external) | 6.00 (2.39) n=27 | 5.37 (1.86) n=27 | 5.81 (2.13) n=27 | 5.59 (2.61) n=27 | 5.59 n=27 |
| Probability (internal) | 6.35 (2.62) n=27 | 5.20 (2.29) n=27 | 6.52 (2.36) n=27 | 6.33 (2.29) n=27 | 6.01 n=27 |
| Probability w/range | 6.46 (2.13) n=27 | 6.04 (2.65) n=27 | 5.93 (2.63) n=27 | 6.96 (1.97) n=27 | 6.31 n=27 |
| Total | 6.27 n=81 | 5.54 n=81 | 6.09 n=81 | 6.30 n=81 | |

^a Mean totals for the between subject condition of probability format are make up of only those observations in the numerical conditions. In other words, the responses in the pure narrative condition were not included in these means because there was no explicit probability information present in this condition. This is necessary because the experimental design is not fully crossed.

Table 21. The effect of stated likelihood and likelihood format on perceived source credibility.

| | Narrative | 1% | 5% | 10% | Total ^a |
|--------------------------|---------------------|---------------------|---------------------|---------------------|--------------------|
| Probability (external) | 5.58 (1.91) n=27 | 4.97 (1.50) n=27 | 5.16 (1.58) n=27 | 5.55 (1.76) n=27 | 5.23 n=27 |
| Probability (confidence) | 5.68 (1.92) n=26 | 5.06 (1.60) n=26 | 5.83 (1.77) n=26 | 6.41 (1.35) n=26 | 5.77 n=26 |
| Probability w/range | 6.85 (1.76) n=27 | 6.33 (1.94) n=27 | 6.44 (1.85) n=27 | 6.89 (1.63) n=27 | 6.55 n=27 |
| Total | 6.04 n=80 | 5.46 n=80 | 5.81 n=80 | 6.28 n=80 | |

^a Mean totals for the between subject condition of probability format are make up of only those observations in the numerical conditions. In other words, the responses in the pure narrative condition were not included in these means because there was no explicit probability information present in this condition. This is necessary because the experimental design is not fully crossed.

Perceived usefulness.

Explicit likelihood and narrative information: The pure narrative and the average of numerical forecasting conditions did not differ in perceived usefulness, $t(318) = 1.145$, $p = .254$. Within the numerical forecasting conditions, there was a significant linear trend across stated likelihood, $t(237) = 2.434$, $p = .016$, $ES = .53$, such that consumers rated forecasts with higher stated likelihood as more useful for decision making. Inspection of the means shows that consumers actually thought the pure narrative report was more valuable than forecasts with 1% or 5% probabilities (averaging across format), and roughly equal in value to forecasts in the 10% condition. This makes sense if consumers were judging the forecasts with lower stated likelihood as more “wrong” in hindsight, and a forecast that is wrong is not as useful as a forecast that says nothing at all about likelihood (pure narrative condition).

In addition to using the stated likelihood information to judge how useful the report would have been for decision making, there was also an association between ratings of the credibility and coherence of the evidence and perceptions of usefulness, $t(225) =$

9.353, $p < .001$, $ES = 2.42$, such that consumers found a forecast to be more useful the greater they perceived the credibility and coherence of the narrative evidence summary to be.

The effect of likelihood format: There were no significant differences due to likelihood format, although there was a trend for the range condition to be rated as more useful than the external point estimate condition, $t(78) = 1.611$, $p = 0.11$, $ES = .25$.

Numeracy: As in Study 2, consumers higher in numeracy rated the forecasts lower in value overall, $t(75) = -2.793$, $p = .007$, $ES = .41$.

Source Credibility.

The results for source credibility were very similar to the results for perceived usefulness.

Explicit likelihood and narrative information: There was a significant linear trend across stated likelihood, $t(236) = 3.976$, $p < .001$, $ES = .76$, such that consumers rated forecasts with higher stated likelihood higher in source credibility. There was also an association between ratings of the credibility and coherence of the evidence and perceptions of usefulness, $t(227) = 9.500$, $p < .001$, $ES = 3.50$, such that consumers found a forecast to have more source credibility the greater they perceived the credibility and coherence of the narrative evidence summary to be.

The effect of likelihood format: There was a significant difference between the range condition and point estimate conditions, $t(78) = 3.463$, $p = .001$, $ES = .55$, such that consumers rated the range condition higher in source credibility than the point estimate conditions.

Numeracy: There were no significant effects of consumer numeracy.

Perceived Likelihood

Specific research questions are detailed below:

1. In hindsight, will the pure narrative forecast elicit higher perceptions of likelihood than the numerical conditions? In other words, will consumers be sensitive to the explicit likelihood estimates in the numerical forecasting conditions, even though they already know the outcome of the forecasted event?
2. Will consumers use the stated likelihood and the properties of the narrative summary to make likelihood judgments in hindsight?
3. Will the format of the likelihood information or consumer numeracy affect perceptions of likelihood in hindsight?

In this experiment, consumers were asked to look back at the intelligence forecast written before the eventual attack and to rate how they would have rated the likelihood of the attack if they had been given this forecast before the attack occurred. Table 22 shows the effect of stated likelihood and likelihood format on perceptions of likelihood in hindsight.

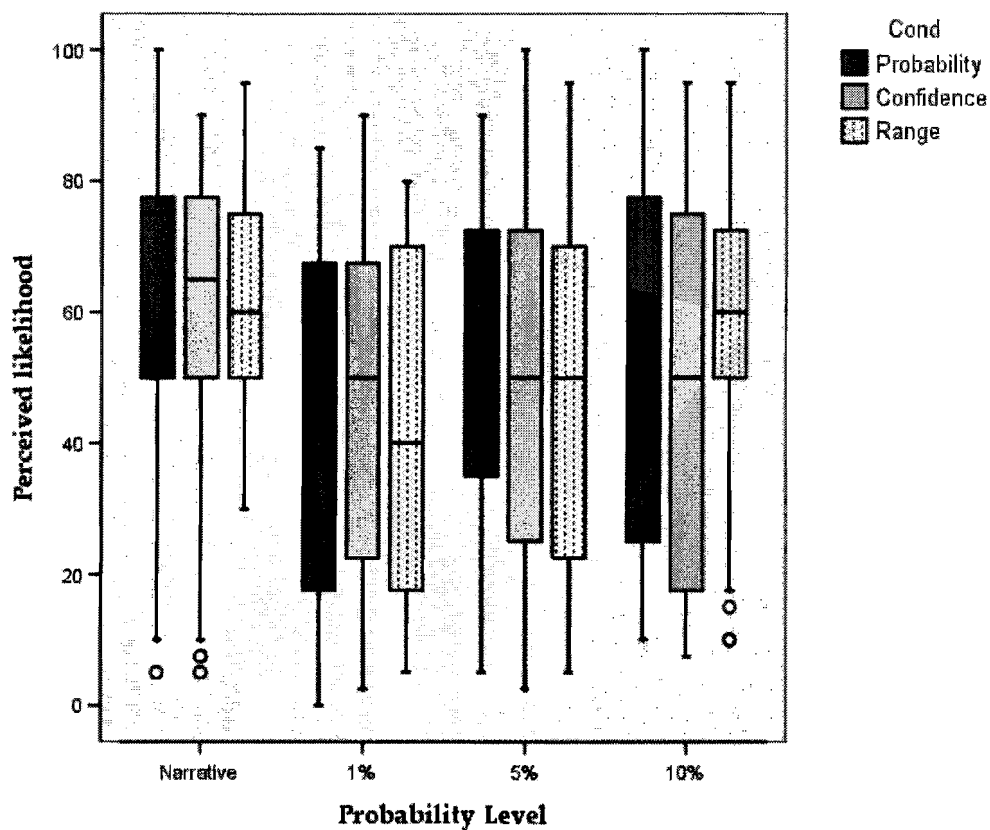
Table 22. The effect of stated likelihood and likelihood format on perceptions of likelihood.

| | Narrative | 1% | 5% | 10% | Total ^a |
|---------------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------|
| Probability (external) | 63.70 (23.40) n=27 | 43.70 (27.76) n=27 | 51.48 (26.52) n=27 | 51.48 (30.47) n=27 | 48.89 n=27 |
| Probability (internal) | 59.54 (25.65) n=27 | 43.80 (27.57) n=27 | 48.80 (30.25) n=27 | 51.39 (30.00) n=27 | 48.00 n=27 |
| Probability w/range | 60.93 (17.65) n=27 | 40.56 (27.15) n=27 | 45.00 (27.98) n=27 | 56.57 (24.41) n=27 | 47.38 n=27 |
| Total | 61.39 n=81 | 42.69 n=81 | 48.43 n=81 | 53.15 n=81 | |

^aMean totals for the between subject condition of probability format are made up of only those observations in the numerical conditions. In other words, the responses in the pure narrative condition were not included in these means because there was no explicit probability information present in this condition. This is necessary because the experimental design is not fully crossed.

Unlike Study 3, where consumers judged the likelihood of potential terrorist plots from these intelligence forecasts without the benefit of hindsight, the distributions of likelihood ratings were not drastically skewed. Figure 17 shows the distributions of hindsight likelihood ratings by stated likelihood and likelihood format.

Figure 17. The distributions of hindsight likelihood ratings by stated likelihood and likelihood format.



Additionally, it is clear from Table 22 that the likelihood ratings are much higher in this experiment than when these same forecasts were judged without hindsight (Study 3). This is an example of a classic hindsight effect, in that even when told to make a judgment with the information that was only available before the outcome of an event, the knowledge of the outcome biased likelihood estimates in the direction of the outcome.

Thus, in this case because the event did occur consumers judged that they would have thought that the attack was very likely. In addition, hindsight likelihood ratings were higher in the narrative condition as opposed to the numerical conditions, $t(318) = 5.568$, $p < .001$, $ES = 1.57$.

In the next set of analyses, hindsight likelihood judgments were modeled as a function of stated likelihood, likelihood format, ratings of the credibility and coherence of the evidence, and consumer numeracy.

Stated likelihood and properties of the narrative summary. As in Study 3, consumers showed sensitivity to the stated likelihood in the forecasts, reporting higher perceptions of likelihood as the stated likelihood increased, $t(231) = 2.798$, $p = .006$, $ES = .49$. In addition, higher ratings of the credibility and coherence of the narrative summary were associated with higher perceived likelihood in hindsight, $t(231) = 7.692$, $p < .001$, $ES = 2.77$. As in Study 3, perceptions of the credibility and coherence of the narrative summary had a larger effect of perceptions of likelihood than the stated likelihood manipulation (1%, 5%, 10%). Unlike Study 3, however, the numeracy level of the consumer did not significantly moderate the use of the evidence properties or the stated probability information, although the effects were in the same direction.

Numeracy. There was a relatively small effect for numeracy, $t(77) = -1.916$, $p = .059$, $ES = .26$, such that consumers lower in numeracy reported higher perceived likelihood.

Summary and Discussion

This study was designed to explore consumer perceptions of intelligence forecasts in hindsight. Specifically, consumers were given the details of a terrorist attack that occurred and then they were asked to examine an intelligence forecast that was produced before the attack. In these situations people have shown what has been called a hindsight bias, in that they tend to overestimate the likelihood or the ease to which the event could have been predicted. In other words, the knowledge about the eventual outcome of the forecasted event biases perceptions of what was known or what could have been predicted before the event occurred. This effect is important because this is exactly the

situation that intelligence analysts may find themselves in if a terrorist attack were to occur, particularly an attack that they had previously assigned a relatively low likelihood in an intelligence forecast. As stated above, this is one of the reasons that analysts may prefer purely narrative forecasts, so as not to look like they are making deterministic predictions. Strictly speaking, a probabilistic forecast can never be wrong because, by definition, the forecaster is not making a deterministic claim about whether an event will occur or not. However, a consumer looking back at a forecast in hindsight (already knowing that the forecasted attack did occur) may perceive forecasts with smaller probabilities as more “wrong” than those with higher probabilities. Previous research discussed in Chapter II suggests that some consumers do tend to look at probabilistic forecasts in just such a deterministic manner.

Pure narrative forecasts versus numerical forecasts

Does presenting a more ambiguous pure narrative forecast reduce perceptions of blame as compared to numerical forecasts when a negative target event occurs? Overall, pure narrative forecasts did not result in significantly lower perceptions of blame than forecasts with explicit estimates of likelihood and potential harm. This was somewhat surprising because the numerical forecasts all had relatively low estimates of likelihood for the event (1%, 5%, 10%). One might expect that the forecaster in the pure narrative forecast would not have been blamed for making a poor forecast because there was no likelihood value on which to base this judgment. However, the mean rating of blame was lower in the pure narrative condition as compared to the 1% and 5% numerical forecasting conditions, and if the likelihoods were estimated to be much lower there would likely be a significant difference between pure narrative and numerical estimates. As will be discussed below, consumers were sensitive to the stated likelihood when assigning blame.

In addition, the pure narrative forecast was perceived to have more usefulness and source credibility than the forecasts with the lowest levels of probability. This is most likely because consumers were sensitive to the likelihood information when assessing the

quality of the forecasts in hindsight and the low estimates of likelihood were perceived to be more “wrong”, and therefore not as valuable or credible.

The effect of both stated likelihood and properties of the narrative summary

The results suggest that participants do tend to assign more blame to a forecaster that assigned a smaller likelihood to a terrorist attack that eventually occurred. Interestingly, this effect was not moderated by numeracy. It appears that consumers at all levels of numerical ability take likelihood into account when judging intelligence forecasts in hindsight. In addition, perceptions of the credibility and coherence of the evidence set were not found to significantly predict perceptions of blame. It appears that consumers focus more on the stated likelihood information and ignore the narrative evidence summary when making judgments of blame. However, consumer perceptions of the potential harm of the attack were also associated with perceived blame, such that consumers that perceived greater potential harm in the forecasted event assigned more blame to the forecaster.

In terms of perceived usefulness and source credibility, consumers found the forecasts with lower stated probabilities and forecasts that were perceived to have a less credible and coherence evidence set to have less usefulness and source credibility.

Effects of likelihood format

Within the numerical probability conditions, the range condition elicited slightly less blame than the point estimate conditions. The increased blame in the point estimate conditions may be due to the fact that they appear more precise, and are therefore perceived as being more “wrong”. In the range condition, by contrast, the analyst is communicating a certain amount of uncertainty in the analysis, and may appear less blameworthy. Finally, the range condition was thought to have more value and source credibility than the point estimate conditions in hindsight.

Previous research suggested that consumers would assign more blame to forecasters that made incorrect forecasts when they expressed likelihood in an internal format

(confidence) as opposed to an external format (Fox and Malle, 1995). No significant differences were found between internal and external likelihood formats.

Perceptions of likelihood in hindsight

Consumers showed clear hindsight effects in their perceptions of the likelihood of the forecasted event after already knowing the outcome. The mean likelihood ratings were substantially higher than the likelihood ratings for the same forecasts in Study 3. Although consumer likelihood judgments were much higher in hindsight, consumers were still sensitive to the stated probability information and perceptions of the credibility and coherence of the evidence set. As in Study 3, perceptions of the narrative summary had a larger effect on perceived likelihood in hindsight than the stated probability manipulation (1%, 5% to 10%). Previous research suggests that when causal or scenario-based information is present in hindsight, this information will be reevaluated in light of the outcome knowledge (see Chapter II: Hindsight Effects). Participants did use the properties of the narrative evidence to make their hindsight likelihood judgments, and a reevaluation of the narrative evidence summary is a likely explanation for the drastically increased perceptions of likelihood in hindsight.

CHAPTER VI

CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH DIRECTIONS

Conclusions and Implications

The main goal of many political and intelligence forecasts is to communicate risk to decision makers. These forecasts should be communicated in a way that effectively transmits risk information from analyst to consumer. However, standard reporting methods in policy and intelligence analysis rarely involve explicit, numerical estimates of uncertainty, even though several experts have argued that the explicit treatment of uncertainty will lead to improved analysis and risk communication. Standard reporting methods for intelligence forecasts most often involve a scenario-based or narrative discussion of the evidence and possible future states of the world, and any numerical estimates of uncertainty would likely accompany this narrative presentation.

The primary purpose of presenting numerical estimates of uncertainty is to communicate, as accurately as possible, the risk estimates generated by the analyst to the intelligence consumer. For example, numerical estimates of likelihood are more precise than narrative descriptions of evidence and it has been presumed that they allow more consistent interpretation by consumers. Although much previous research has focused on the analytic techniques that can be used to estimate these numerical quantities, how these analytic results should be reported for the benefit of consumers has received less attention. The work in this dissertation has focused on risk communication in intelligence forecasts from the consumer's perspective.

Perceptions of Intelligence Forecasts with Numerical Likelihood and Narrative Information

The intelligence consumer is faced with a difficult task because both the numerical estimates of uncertainty and the narrative supporting evidence could be used to inform

perceived likelihood and risk. In the present studies, consumers did perceive forecasts with explicit likelihoods as more useful than forecasts with only a narrative evidence summary. However, the majority of consumers did not consistently use stated likelihood to inform their perceptions of likelihood, and the properties of the narrative summary had a strong influence on perceptions of likelihood. These results shows that even “precise” explicit statements of likelihood are not necessarily evaluable by consumers and the perception of likelihood is affected by the contextual information available to the judge (see Windschit et al., 1999, 2002; Hsee, 1995; Hendrickx et al., 1989, 1992; Yates et al., 1996).

One of the reasons that analysts may be reluctant to assign numerical likelihood estimates to forecasts is that they feel they may be blamed if a relatively small likelihood is attached to an event that eventually occurs. Strictly speaking, a probabilistic forecast can never be wrong because, by definition, the forecaster is not making a deterministic claim about whether an event will occur or not. Looking at a series of forecasts is the only way to assess the skill or calibration of a forecaster. Some consumers, however, did tend to think of these single-event forecasts in a deterministic manner by assigning more blame to a forecaster who assigned a smaller likelihood to a terrorist attack that eventually occurred. When evaluating forecasts in hindsight consumers were found to be sensitive to stated likelihood but not the properties of the narrative evidence summary (although perceptions of harm were also predictive of blame judgments). These results suggest that at least some of the consumers in the sample did not fully appreciate the nature of probabilistic statements. These consumers may perceive these statements as ratings of event propensity, with probabilities above 50% being correct and those below 50% being incorrect and probabilities further away from the correct side of the distribution as more “wrong” (e.g. 10% is more “wrong” than 30%). Unfortunately, this is exactly the type of hindsight interpretation of probabilistic estimates that analysts may fear. It is unclear how best to deal with this issue, although it is possible that simple educational interventions focused on the nature of probabilistic statements could help (see discussion of Numeracy below).

In addition, there did appear to be benefits of presenting numerical likelihood estimates in these forecasts. Consumer's sensitivity to the stated likelihood helped to control hindsight likelihood judgments as compared to narrative-only forecasts. Consumer judgments of the likelihood of an attack in hindsight were much higher when presented with narrative-only forecasts as compared to forecasts accompanied by a relatively low stated likelihood.

Implications

Reporting explicit estimates of uncertainty in a forecast does not necessarily mean that this information will be consistently interpreted or used by consumers of the forecast, particularly when presented with supporting narrative evidence. Consumers may more consistently use the numerical estimates to inform their perceptions of risk and likelihood if supporting information is not presented (see results from Study 2), but it is unlikely that consumers would trust or feel comfortable using a purely numerical forecast in this domain. If an analyst presented a report consisting of only numerical estimates of likelihood and potential harm, the consumer would most likely want to know on what basis the analyst came to that conclusion. For example, Yates et al. (1996) reported several experiments in which consumers evaluated forecasts concerning the outcome of lawsuits. In these experiments consumers were presented with only numerical forecasts. However, consumers often expressed an interest in having more justification about the methods and evidence that the forecasters used to make their judgments. Yates et al (1996) note that "... it should be irrelevant how a consultant arrives at his or her assessments, only that those judgments are reliably good in a statistical sense. But that is apparently not good enough for many consumers." (pg. 45). Forecasting events in the domain of politics and human affairs may not be perceived in the same way as the engineer reporting the likelihood of an in flight engine failure on the commercial airliner. Consumers may be far more likely to take the numerical estimate of the engineer at face value, and not press him or her for details about how this estimate was generated. In the political and intelligence domain, however, consumers may intuitively understand the

difficulties and uncertainties involved in forecasting human events, and they may not be likely to take a probability point estimate at face value without inquiring about the evidence behind this judgment.

Consumers willingness to accept numerical forecasts at face value may be limited to forecasting situations where statistical information of past performance or deterministic laws governing the phenomenon are perceived to be important to the estimation of the likelihood. In forecasting situations that are perceived to be based on the examination of evidence and reasoning processes such as analogy and scenario generation, consumers will most likely want to see supporting evidence (Yates, 1996). The present results suggest that any supporting information presented to consumers may have a large impact on perceptions of the likelihood and risk of the event, potentially overwhelming, or at least greatly affecting, the numerical likelihood estimates that are generated by the forecaster. This may result from the fact that the layperson is well practiced in “sense making” and reasoning processes based on scenario generation and the examination of evidence, and these consumers may automatically engage in this type of reasoning when presented with narrative evidence-based information. In contrast, if the engineer described above presented the technical information about engine reliability, these common reasoning processes would not be clearly applicable, and the non-expert consumer would most likely use the engineer’s likelihood estimate.

The supporting evidence underlying a forecast been shown to have a large impact on consumer perceptions of risk and likelihood. Thus, forecasters must be extremely careful in choosing the types of information that are reported and the format of that information, even when numerical estimates of likelihood or risk are reported as well. Ideally, the explicit estimates of likelihood and potential harm would work in concert with any supporting narrative information, providing the consumer with a complete picture of the risk associated with the forecasted event.

The Formatting of Numerical Likelihood in Intelligence Forecasts

The format of the stated likelihood information moderated the impact of stated likelihood on perceived risk, likelihood and perceptions of the “quality” of the report. The use of verbal probability estimates was found to be a poor method of transferring likelihood information from analyst to consumer. Consumers judged forecasts with verbal probability estimates to be less useful and the forecaster that reported them as less knowledgeable and less trustworthy.

Among the point estimate numerical formats, consumers were more consistent in using stated likelihood to inform perceived likelihood in the percentage format as compared to the frequency format. Previous research has focused on the benefits of frequency information over single-event probability formats (i.e. percentage and decimal formats; see Chapter II), but this may be restricted to situations when statistical reasoning is involved. In the forecasting situation, the likelihood estimate is only meant to transmit information to the consumer and it appears that the frequency format may actually be more confusing. When likelihood is represented as a ratio (i.e. frequency format), both the numerator and the denominator must be evaluated in relation to one another, while in the percentage format there is just a single number that needs to be evaluated. In addition, the representation of likelihood as a relative frequency may not be readily understood when it is attached to a single, non-repeating event.

Performing sensitivity analysis and reporting a range of plausible parameter estimates in an important topic in risk communication and forecasting. This is mainly because of the complexity present in many policy/intelligence domains and the sensitivity of the results to changes in the initial conditions and inputs. Ranges of plausible values are also useful for reporting second-order uncertainty to consumers and reducing the perceived precision of these estimates that results when only point estimates are presented.

Risk and likelihood perceptions were not as consistent in the range condition (i.e. higher risk and likelihood perceptions for higher stated likelihood ranges) as they were in the point estimate conditions. The range of likelihood estimates allows more flexibility in the interpretation of the estimate (i.e. Does one focus on the best estimate or the low or

high end of the range?), and likely decreases the consistency with which consumers use this information. However, consumers found the range format to be more useful and the forecaster more knowledgeable and trustworthy than the point estimate format, but only at higher stated likelihood (i.e. 20% versus 5%). Consumers were also less affected by the narrative evidence summary when judging likelihood in the percentage with range condition. This result was unexpected. Previous research suggests that when participants are presented with a range of values, they will tend to ignore the information and use other more easily evaluated information to make the judgment at hand (Hsee, 1995). Thus, one might expect that consumers would be more influenced by the narrative information in the range condition as compared to the point estimates conditions. It is unclear why the opposite effect was observed here, but it suggests that presenting a range of values will not necessarily force consumers to focus on other information to make a judgment or decision.

In addition, when evaluating an intelligence forecast in hindsight, consumers assigned lower levels of blame to forecasters when they presented their forecasts with a range of estimates. Consumers also rated the range format higher in usefulness and source credibility. The increased blame in the point estimate conditions may be due to the fact that they appear more precise, and are therefore perceived as being more “wrong”. In the range condition, by contrast, the analyst is communicating a certain amount of uncertainty in the analysis, and may appear less blameworthy

Implications

As discussed by many researchers, expressing uncertainty in verbal form is not likely to be an effective method of communication between analyst and consumer, at least not without some kind of reference scale accompanying the forecast (and at that point one might as well use a numerical scale). In addition, representing likelihood as a single-event probability (e.g. in percentage form) appears to be a better choice than a frequency representation.

These results suggest that presenting likelihood as a range has both positive and negative repercussions. Consumers may not clearly differentiate forecasts as well when a range of likelihood values is reported, although they seem to be less affected by the supporting narrative information. In addition, less blame is assigned in hindsight when a range of values is reported. Consumers also find range formats to be more useful and believe the forecaster is more knowledgeable and trustworthy, at least at higher probabilities.

Presenting ranges and confidence intervals may turn out to be the only plausible method of quantitative forecasting in the political and intelligence domains. Analysts are unlikely to be comfortable reporting point estimates in many situations, both because of the complexity of the problems and the insufficient data on which these judgments are often based, and that they do not want consumers to perceive these estimates as “precise”. Consumers will also benefit from the additional information provided by confidence ranges. Schrage (2005) notes this as one of the important advantages to reporting uncertainty in intelligence forecasts, in that consumers will have more information about the confidence that a forecaster has in his or her conclusions. If the reporting of confidence ranges becomes standard practice in intelligence forecasting, additional research will be needed to more fully understand the positive and negative effects that this approach will have on consumers.

Individual Differences in the Numerical Ability of Consumers

Individual differences in numerical ability also had an effect on how consumers perceived and used quantitative forecasts. Consumers lower in numeracy focused more on the properties of the narrative summary and did not use the stated likelihood to inform their perceptions of risk as much as higher numerate consumers. In addition, consumers with different levels of numeracy also perceived particular likelihood formats to be more useful for decision making, and found the forecaster that reported these likelihoods to be more knowledgeable and trustworthy.

These results add to a series of recent findings that connect differences in numerical ability, or how well people can evaluate and use numbers, to judgment and decision making behavior (see Dieckmann, 2007 for a review). Whenever a judgment or decision making task involves the evaluation of numbers, consumers may choose very different reasoning strategies for completing these tasks depending on numerical ability. This effect will likely be magnified when there are other sources of information, beside the numbers, that may be more easily used to make the judgment or decision at hand. Intelligence forecasts with both explicit numerical information and a narrative evidence summary are an excellent example of just such a situation.

Implications

Consumers of political and intelligence forecasts will vary in their comfort with numbers and their ability to use and evaluate numerical information. These differences in numeracy may greatly affect how consumers view the conclusions of the forecast and how well numerical information (in this case probabilistic information) can be used to transfer risk information from analyst to consumer. One way to alleviate this problem is for forecasters to find methods of reporting numerical information that is evaluable to consumers at all levels of numerical ability. The results from several recent studies suggest that alternative presentations of health-related information may make this information more evaluable for consumers lower in numerical ability (Peters, Dieckmann, et al., 2007; Peters, Dieckmann, Vastjall, Mertz, et al., 2006). For example, simplifying information displays to ease the cognitive burden of a task and providing verbal category labels to facilitate the evaluability of numerical information have been shown to improve judgment and choice. Methods similar to these or new methods could be developed to help consumers of intelligence forecasts evaluate and understand quantitative forecasts.

A second way to address this problem is to teach consumers about the evaluation and interpretation of any numerical quantities presented in a forecast. For example, this could take the form of short written tutorials describing the interpretation and suggested use of

any numerical information. It is unclear, however, how effective short tutorials will be in improving the understanding of probabilistic information and eventually improving the use of this information by those lower in numeracy.

If there is a question about how any numerical information will be understood by consumers, a forecaster should consider alternative formats for quantitative information to improve evaluability. The forecaster may also consider including a short tutorial describing the interpretation of any numerical information included in the forecast. However, additional research is needed to assess the effectiveness of these interventions.

Limitations

Each of the experiments in this study used simulated intelligence forecasts of potential terrorist plots involving explosive devices in the United States. As discussed in Chapter III, the characteristics of the particular hazard under study will affect laypersons' perceptions of risk (e.g. the controllability or dread risk of the hazard; Slovic, 1987). Thus, it is possible that the results described in this dissertation are in some part restricted to hazards relating to terrorism. Ideally, a representative sample of hazards from the intelligence domain could be tested in future studies to show the generality of the effects that have been described. It may even be possible to use real, unclassified intelligence reports from US government archives.

The sample of research participants may also limit the generalization of these results. Real consumers of intelligence forecasts may have specialized knowledge and backgrounds that may make them respond differently to intelligence forecasts. The recruitment of more educated participants for studies 3 and 4 was done to simulate the likely education level of real intelligence consumers, although there are clearly other contextual factors that were not simulated in the current studies (e.g. worldview, political pressure that may affect a consumer's perception of a forecast, etc).

Another limitation to the generalizability of the results was the relatively narrow range of numerical likelihood that was manipulated (1%-20%). The extent to which consumers are sensitive to the numerical likelihood information and narrative summary

information may depend on the range of likelihood values presented. For example, probability neglect maybe a particularly important problem whenever small likelihoods are communicated in a forecast (e.g. on the order of 1/1000 to 1/1000000; see Sunstein, 2003 for a discussion of probability neglect). For example, consumers may ignore the likelihood estimates because they are too difficult to understand and focus on other information in making their likelihood judgments (e.g. the narrative evidence). More research is needed on how consumers make judgments of likelihood when presented with a wider spectrum of explicit likelihood values.

Finally, the findings relating perceived likelihood to perceptions of the credibility and coherence of the narrative evidence summary are purely correlational in nature. Thus, one should be cautious in any causal interpretation of these findings. For example, it is not clear that perceptions of the coherence and credibility of the evidence set actually lead to greater perceptions of likelihood, or if the increased perceptions of likelihood lead to greater perceptions of the credibility and coherence of the evidence. Ideally, the properties of the narrative evidence summary could be experimentally manipulated to provide a more rigorous test of the causal relationship between these constructs.

Future Research Directions

There are several potentially fruitful future research directions focused on risk communication and intelligence forecasting from the perspective of consumers. These recommendations are based both on the experimental results presented above and a review of the literature in forecasting, risk communication, and intelligence analysis.

There are several different levels of uncertainty that are present in the risk analysis and intelligence forecasting domains. When analyzing a particular problem or set of events there may be uncertainty about the quality or credibility of the evidence, uncertainty about the structural model of situation (how the evidence fits together), uncertainty about the likelihood that particular events will occur in the future, and uncertainty about the potential harm that would result if these events occur (see Chapter

II). Ideally, the forecaster accounts for these different types of uncertainty during the analytic process. It is an open question, however, whether consumers should also be presented with estimates of uncertainty at these different levels. Would consumers be able to interpret this information, and would it improve the judgments and decisions that are eventually based on these forecasts? The results of the present experiments begin to address how consumers would respond to uncertainty relating to the likelihood of future events, with or without second order uncertainty around these estimates. Although some authors suggest presenting additional levels of uncertainty in intelligence forecasts (e.g. Schrage, 2005), future work should explore how feasible it will be for analysts to estimate this uncertainty, and how well consumers could use this information when interpreting a forecast and making subsequent judgments and decisions.

Although single event forecasts are likely to be reported to consumers in the intelligence community, there are also situations in which consumers will need to be informed about numerous potential threats simultaneously. Ideally, these threats could be reported in a format that facilitates trade-offs and comparisons among them. Future research could be aimed at identifying the optimal methods of presenting multiple potential threats simultaneously (see Horowitz & Haines, 2003). Researchers and practitioners should be sensitive to the psychological limitations of consumers who will need to understand and make use of these forecasts.

As long as quantitative intelligence forecasts are accompanied by a narrative evidence summary, consumer's perceptions of the risk and quality of forecasts will be greatly dependent on the nature of this supporting information. Future research should focus on the specific characteristics of this supporting information that affect consumer perceptions of risk and quality. It will also be very important to explore how the characteristics of the supporting information interact with explicit estimates of uncertainty. For example, perceptions of the coherence and credibility of the narrative evidence summary were found to be predictive of consumer perceptions. These characteristics will need to be studied in more rigorous experimental designs in the future, and there are several other potentially important characteristics that may affect

consumer perceptions. For example, the completeness of the explanation, the presence of alternative explanations (either implicit or explicit), and the vividness of the description are all interesting factors of the supporting information that may have strong effects of how a forecast is perceived by consumers.

Future research should also focus on ways of presenting intelligence forecasts that makes them interpretable to consumers with a range of numerical abilities. Probabilistic forecasts are likely to be lost on consumers who do not have the basic numerical skills to interpret the uncertainty information presented by the forecaster. As discussed above, this research could focus on the ways of making numerical information more evaluable to consumers or on ways of teaching consumers about the meaning and interpretation of the numerical information presented.

Hindsight effects in the intelligence forecasting domain is also a very interesting research direction considering the recent high profile intelligence “failures” and the intelligence reports and forecasts that are now being scrutinized after the fact. Future research should further explore how both the quantitative and qualitative properties of intelligence forecasts affect judgments in hindsight. Ideally, future research will identify the types of information that should be included and specific formats for intelligence forecasts that minimize hindsight effects.

APPENDIX A

SCENARIOS TESTED IN PRELIMINARY STUDY 1

Note: Below are the four intelligence scenarios that were tested in Preliminary Study 1. In addition, one example of an actual Presidential Daily Briefing (PDB) is also included (“Bin Laden Determined to Strike in US”). The simulated intelligence forecasts were roughly modeled after historical PDBs and other intelligence reports available in the public domain.

Intelligence Report #1

Intelligence Report:

Yesterday afternoon a foreign newspaper printed a statement from the militant group XXX warning of an attack on the US.

Four months ago, an informant warned that the militant group XXX had tried to purchase a quantity of an unknown explosive. Whether they succeeded in purchasing the explosives is unknown.

The FBI intercepted a cellular telephone call between individuals with suspected links to the militant group XXX. Washington, DC was mentioned repeatedly in the conversation, although they did not reveal any information about an impending attack. The call was intercepted last week and originated within the US.

The FBI has also reported suspicious activity consistent with the surveillance of federal buildings in Washington, DC. This activity has been observed on numerous occasions over the last several months.

The militant group XXX has used explosives against government buildings in foreign countries in the past.

Intelligence Report #2

Intelligence Report:

A tip from an anonymous informant recently led the FBI to the apartment of two men suspected of working for the militant group YYY. When the FBI arrived the men had already left, but investigators did discover simple maps and timetables of the railway systems in Chicago, IL.

Several months ago, a videotaped statement by the leader of the militant group YYY appeared on the Internet. Among other things, the leader alluded to a recent train bombing in Portugal and warned that the United States would be next.

Three months ago, analysts doing routine satellite monitoring of a known YYY training camp reported an increase in activity. It appeared that members of YYY were experimenting with explosive devices.

The YYY militant group has been implicated in several train bombings over the last several years. The most recent attack in Portugal was powerful enough to completely destroy one train car filled with passengers and completely derail the train.

Both the FBI and the Chicago Police have reported suspicious activity around train stops in the city. This activity has been observed on numerous occasions over the last several months.

Intelligence Report #3

Intelligence Report:

On a few different occasions port authorities have stopped and questioned pairs of men trespassing in New York City ports. Each time the men were in the areas of the port where passenger ships dock.

A few weeks ago, a website with ties to the militant group ZZZ posted a statement that warned of attacks on the US. It specifically mentioned that the next attacks would be aimed at a vulnerable place, since so much security has been focused on air travel.

On a tip from an undercover agent, the FBI recently captured a wanted member of the militant group ZZZ. He revealed that group leaders had discussed attacking a port in New York City. He claimed to not know of any details concerning an attack and seemed unsure that members of the group had acquired the necessary explosives.

The FBI has also bugged the apartment of two suspected members of ZZZ. The men have been overheard discussing the technical details of previous terrorist attacks, as well as discussing preparations for leaving the city in the near future.

The ZZZ militant group has used explosives to attack targets in the past.

Intelligence Report #4

Intelligence Report:

Several national security experts have predicted a terrorist attack during a large sporting event in US. The high concentration of people in a relatively small area is the obvious draw of this type of attack.

A member of the militant group VVV was recently apprehended abroad. He revealed that the leadership of VVV had discussed several different plans to use explosives in the US. One plan was to coordinate several simultaneous explosive attacks in a highly populated area. Members of VVV have carried out attacks of this nature before.

In the last several months, both local authorities and the FBI have increased surveillance of professional basketball, baseball, football, and hockey events in the Los Angeles area. On one occasion, a suspicious package was left in a crowded area at a professional basketball game. The package turned out to be a hoax, but several authorities reported suspicious persons possibly observing the response. There is no way to be sure, but the hoax package could have been used to test the response of security and law enforcement.

Last week, the FBI confiscated financial statements and froze the bank accounts of a Los Angeles lawyer suspected of partially supporting members of VVV in the US. In the financial statements were records of a recent purchase of "military materials". It is unknown what exactly was purchased or where the materials are located now.

Declassified and Approved
for Release, 10 April 2004

Bin Ladin Determined To Strike in US



Clandestine, foreign government, and media reports indicate Bin Ladin since 1997 has wanted to conduct terrorist attacks in the US. Bin Ladin implied in US television interviews in 1997 and 1998 that his followers would follow the example of World Trade Center bomber Ramzi Yousef and "bring the fighting to America."

After US missile strikes on his base in Afghanistan in 1998, Bin Ladin told followers he wanted to retaliate in Washington, according to a [REDACTED] service.

An Egyptian Islamic Jihad (EIJ) operative told an [REDACTED] service at the same time that Bin Ladin was planning to exploit the operative's access to the US to mount a terrorist strike.

The millennium plotting in Canada in 1999 may have been part of Bin Ladin's first serious attempt to implement a terrorist strike in the US. Convicted plotter Ahmed Ressaam has told the FBI that he conceived the idea to attack Los Angeles International Airport himself, but that Bin Ladin lieutenant Abu Zubaydah encouraged him and helped facilitate the operation. Ressaam also said that in 1998 Abu Zubaydah was planning his own US attack.

Ressaam says Bin Ladin was aware of the Los Angeles operation.

Although Bin Ladin has not succeeded, his attacks against the US Embassies in Kenya and Tanzania in 1998 demonstrate that he prepares operations years in advance and is not deterred by setbacks. Bin Ladin associates surveilled our Embassies in Nairobi and Dar es Salaam as early as 1993, and some members of the Nairobi cell planning the bombings were arrested and deported in 1997.

Al-Qa'ida members—including some who are US citizens—have resided in or traveled to the US for years, and the group apparently maintains a support structure that could aid attacks. Two al-Qa'ida members found guilty in the conspiracy to bomb our Embassies in East Africa were US citizens, and a senior EIJ member lived in California in the mid-1990s.

A clandestine source said in 1998 that a Bin Ladin cell in New York was recruiting Muslim-American youth for attacks.

We have not been able to corroborate some of the more sensational threat reporting, such as that from a [REDACTED] service in 1998 saying that Bin Ladin wanted to hijack a US aircraft to gain the release of "Blind Shaykh" Umar 'Abd al-Rahman and other US-held extremists.

continued

For the President Only
6 August 2001

[REDACTED]


Declassified and Approved
for Release, 10 April 2004

Declassified and Approved
for Release, 10 April 2004

— Nevertheless, FBI information since that time indicates patterns of suspicious activity in this country consistent with preparations for hijackings or other types of attacks, including recent surveillance of federal buildings in New York.

The FBI is conducting approximately 70 full field investigations throughout the US that it considers Bin Ladin-related. CIA and the FBI are investigating a call to our Embassy in the UAE in May saying that a group of Bin Ladin supporters was in the US planning attacks with explosives.

For the President Only
6 August 2001

 Declassified and Approved
for Release, 10 April 2004

APPENDIX B

MATERIALS FOR STUDY 2

Note: Because of the fully between subjects design, there were 16 different experimental conditions. The two scenarios below are the summary only and summary with narrative evidence conditions in the verbal probability condition for the low level of probability. The additional manipulations of probability level and probability format are displayed in brackets. Only the information in the sentence in bold was manipulated across the probability level and format conditions. The dependent variables and the numeracy measure used in Study 2 follow the scenarios.

Evaluating an Intelligence Report

INSTRUCTIONS: Imagine that you receive the following intelligence report about a possible terrorist attack. Read the report carefully. On the next page you will make a series of judgments about this report.

Intelligence Report:

The militant group XXX might use explosives to attack a federal building in Washington, DC. If the attack occurs, a plausible worst-case scenario would be 1000 deaths and injuries and 50 million dollars in property damage.

Based on the evidence outlined above and our professional judgment and experience, we estimate that this attack is highly unlikely over the next six months.

High verbal: [... **we estimate that this attack is fairly unlikely over the next six months.**]

Low percentage: [... **we estimate that the probability that this attack will occur over the next six months is 5%.**]

High percentage: [... **we estimate that the probability that this attack will occur over the next six months is 20%.**]

Low frequency: [... **we estimate that the probability that this attack will occur over the next six months is 5 out of 100.**]

High frequency: [... **we estimate that the probability that this attack will occur over the next six months is 20 out of 100.**]

Low range: [... **our best estimate of the probability that this attack will occur over the next six months is 5%, but the probability could be as low as 1% or as high as 10%.**]

High range: [... **our best estimate of the probability that this attack will occur over the next six months is 20%, but the probability could be as low as 10% or as high as 30%.**]

Evaluating an Intelligence Report

INSTRUCTIONS: Imagine that you receive the following intelligence report about a possible terrorist attack. Read the report carefully. On the following pages you will make a series of judgments about this report.

Intelligence Report:

Yesterday afternoon a foreign newspaper printed a statement from the militant group XXX warning of an attack on the US.

Four months ago, an informant warned that the militant group XXX had tried to purchase a quantity of an unknown explosive. Whether they succeeded in purchasing the explosives is unknown.

The FBI intercepted a cellular telephone call between individuals with suspected links to the militant group XXX. Washington, DC was mentioned repeatedly in the conversation, although they did not reveal any information about an impending attack. The call was intercepted last week and originated within the US.

The FBI has also reported suspicious activity consistent with the surveillance of federal buildings in Washington, DC. This activity has been observed on numerous occasions over the last several months.

The militant group XXX has used explosives against government buildings in foreign countries in the past.

Summary

The militant group XXX might use explosives to attack a federal building in Washington, DC. If the attack occurs, a plausible worst-case scenario would be 1000 deaths and injuries and 50 million dollars in property damage.

Based on the evidence outlined above and our professional judgment and experience, we estimate that this attack is highly unlikely over the next six months.

[The same manipulations outlined above were applied to this condition].

Questions about the Intelligence Report

INSTRUCTIONS: Please answer the following questions about the intelligence report on the previous page.

1. How would you rate the risk associated with this possible attack?

| | | | | | | | | | | |
|------------------|---|---|---|---|------------------|---|---|---|---|----------------------|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Very low Risk | | | | | Moderate Risk | | | | | Very high Risk |

2. How valuable is this intelligence report? In other words, does it provide useful information for determining future actions to take?

| | | | | | | | | | | |
|------------------------|---|---|---|---|--------------------|---|---|---|---|-----------------------|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Not at all valuable | | | | | Fairly valuable | | | | | Extremely valuable |

3. How knowledgeable does this analyst seem about this potential attack?

| | | | | | | | | | | |
|-----------------------------|---|---|---|---|-------------------------|---|---|---|---|----------------------------|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Not at all knowledgeable | | | | | Fairly knowledgeable | | | | | Extremely knowledgeable |

4. How much do you trust that this analyst is giving you complete and unbiased information/conclusions about this potential attack?

| | | | | | | | | | | |
|----------------------|---|---|---|---|-------------------|---|---|---|---|-----------------------|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Very little Trust | | | | | Moderate Trust | | | | | Very high Trust |

[Numeracy Measure]

NUMBERS - you may not use a calculator for any of these questions.

1. Imagine that we roll a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up as an even number?

Answer: _____

2. In the BIG BUCKS LOTTERY, the chances of winning a \$10.00 prize are 1%. What is your best guess about how many people would win a \$10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS?

Answer: _____ people

3. In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car?

Answer: _____ %

4. Which of the following numbers represents the biggest risk of getting a disease?

___ 1 in 100 ___ 1 in 1000 ___ 1 in 10

5. Which of the following numbers represents the biggest risk of getting a disease?

___ 1% ___ 10% ___ 5%

6. If Person A's risk of getting a disease is 1% in ten years, and Person B's risk is double that of A's, what is B's risk?

Answer: _____ % in _____ years

7. If Person A's chance of getting a disease is 1 in 100 in ten years, and person B's risk is double that of A, what is B's risk?

Answer: _____ in _____ years

8. If the chance of getting a disease is 10%, how many people would be expected to get the disease:

A: Out of 100? Answer: _____ people

B: Out of 1000? Answer: _____ people

9. If the chance of getting a disease is 20 out of 100, this would be the same as having a _____% chance of getting the disease.

10. The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?

Answer: _____ people

11. Which of the following numbers represents the biggest risk of getting a disease?

___ 1 chance in 12 ___ 1 chance in 37

12. Suppose you have a close friend who has a lump in her breast and must have a mammography. Of 100 women like her, 10 of them actually have a malignant tumor and 90 of them do not. Of the 10 women who actually have a tumor, the mammography indicates correctly that 9 of them have a tumor and indicates incorrectly that 1 of them does not have a tumor. Of the 90 women who do not have a tumor, the mammography indicates correctly that 81 of them do not have a tumor and indicates incorrectly that 9 of them do have a tumor. The table below summarizes all of this information. Imagine that your friend tests positive (as if she had a tumor), what is the likelihood that she actually has a tumor?

| | Tested positive | Tested negative | Totals |
|-----------------------|-----------------|-----------------|--------|
| Actually has a tumor | 9 | 1 | 10 |
| Does not have a tumor | 9 | 81 | 90 |
| Totals | 18 | 82 | 100 |

Answer: _____

13. Imagine that you are taking a class and your chances of being asked a question in class are 1% during the first week of class and double each week thereafter (i.e., you would have a 2% chance in Week 2, a 4% chance in Week 3, an 8% chance in Week 4). What is the probability that you will be asked a question in class during Week 7?

Answer: _____%

14. Suppose that 1 out of every 10,000 doctors in a certain region is infected with the SARS virus; in the same region, 20 out of every 100 people in a particular at-risk population also are infected with the virus. A test for the virus gives a positive result in 99% of those who are infected and in 1% of those who are not infected. A randomly selected doctor and a randomly selected person in the at-risk population in this region both test positive for the disease. Who is more likely to actually have the disease?

___ They both tested positive for SARS and therefore are equally likely to have the disease

___ They both tested positive for SARS, and the doctor is more likely to have the disease

___ They both tested positive for SARS and the person in the at-risk population is more likely to have the disease.

APPENDIX C

MATERIALS FOR STUDY 3

Note: In Study 3, consumers responded to four different scenarios, which were randomly assigned to one level of the probability level within subject factor (i.e. narrative, 1%, 5%, 10%). Participants were also randomly assigned to one of three levels of the probability format (i.e. probability, confidence, probability w/range). Below are the four different scenarios used in Study 3 at each level of the probability factor. In the actual experiment each scenario was matched with each level of probability in a counterbalanced design. The different levels of the between subject factor of probability format are displayed in brackets. The dependent variables for Study 3 follow the experimental materials. Also, the same numeracy measure used in Study 1 was used in Studies 3 and 4.

Intelligence Report #1

INSTRUCTIONS: Imagine that you receive the following intelligence report about a possible terrorist attack. Read the report carefully. On the following pages you will make a series of judgments about this report.

Intelligence Report:

Yesterday afternoon a foreign newspaper printed a statement from the militant group XXX warning of an attack on the US.

Four months ago, an informant warned that the militant group XXX had tried to purchase a quantity of an unknown explosive. Whether they succeeded in purchasing the explosives is unknown.

The FBI intercepted a cellular telephone call between individuals with suspected links to the militant group XXX. Washington, DC was mentioned repeatedly in the conversation, although they did not reveal any information about an impending attack. The call was intercepted last week and originated within the US.

The FBI has also reported suspicious activity consistent with the surveillance of federal buildings in Washington, DC. This activity has been observed on numerous occasions over the last several months.

The militant group XXX has used explosives against government buildings in foreign countries in the past.

Summary

The militant group XXX might use explosives to attack a federal building in Washington, DC.

[The pure narrative forecast condition was identical at each level of the probability format factor.]

Intelligence Report #2

INSTRUCTIONS: Imagine that you receive the following intelligence report about a possible terrorist attack. Read the report carefully. On the following pages you will make a series of judgments about this report.

Intelligence Report:

A tip from an anonymous informant recently led the FBI to the apartment of two men suspected of working for the militant group YYY. When the FBI arrived the men had already left, but investigators did discover simple maps and timetables of the railway systems in Chicago, IL.

Several months ago, a videotaped statement by the leader of the militant group YYY appeared on the Internet. Among other things, the leader alluded to a recent train bombing in Portugal and warned that the United States would be next.

Three months ago, analysts doing routine satellite monitoring of a known YYY training camp reported an increase in activity. It appeared that members of YYY were experimenting with explosive devices.

The YYY militant group has been implicated in several train bombings over the last several years. The most recent attack in Portugal was powerful enough to completely destroy one train car filled with passengers and completely derail the train.

Both the FBI and the Chicago Police have reported suspicious activity around train stops in the city. This activity has been observed on numerous occasions over the last several months.

Summary

The militant group YYY might use explosives to attack a train in Chicago. If the attack occurs, a plausible worst-case scenario would be 1000 deaths and injuries and 50 million dollars in property damage.

Based on the evidence outlined above and our professional judgment and experience, we estimate that the probability that this attack will occur over the next six months is 1%

Confidence condition: [...we are 1% sure that this attack will occur over the next six months.]

Range condition: [... our best estimate of the probability that this attack will occur over the next six months is 1%, but the probability could be as low as .1% or as high as 5%.]

Intelligence Report #3

INSTRUCTIONS: Imagine that you receive the following intelligence report about a possible terrorist attack. Read the report carefully. On the following pages you will make a series of judgments about this report.

Intelligence Report:

On a few different occasions port authorities have stopped and questioned pairs of men trespassing in New York City ports. Each time the men were in the areas of the port where passenger ships dock.

A few weeks ago, a website with ties to the militant group ZZZ posted a statement that warned of attacks on the US. It specifically mentioned that the next attacks would be aimed at a vulnerable place, since so much security has been focused on air travel.

On a tip from an undercover agent, the FBI recently captured a wanted member of the militant group ZZZ. He revealed that group leaders had discussed attacking a port in New York City. He claimed to not know of any details concerning an attack and seemed unsure that members of the group had acquired the necessary explosives.

The FBI has also bugged the apartment of two suspected members of ZZZ. The men have been overheard discussing the technical details of previous terrorist attacks, as well as discussing preparations for leaving the city in the near future.

The ZZZ militant group has used explosives to attack targets in the past.

Summary

The militant group ZZZ might use explosives to attack a passenger ship in New York City. If the attack occurs, a plausible worst-case scenario would be 1000 deaths and injuries and 50 million dollars in property damage.

Based on the evidence outlined above and our professional judgment and experience, we estimate that the probability that this attack will occur over the next six months is 5%

Confidence condition: [...we are 5% sure that this attack will occur over the next six months.]

Range condition: [... our best estimate of the probability that this attack will occur over the next six months is 5%, but the probability could be as low as .5% or as high as 10%.]

Intelligence Report #4

INSTRUCTIONS: Imagine that you receive the following intelligence report about a possible terrorist attack. Read the report carefully. On the following pages you will make a series of judgments about this report.

Intelligence Report:

Several national security experts have predicted a terrorist attack during a large sporting event in US. The high concentration of people in a relatively small area is the obvious draw of this type of attack.

A member of the militant group VVV was recently apprehended abroad. He revealed that the leadership of VVV had discussed several different plans to use explosives in the US. One plan was to coordinate several simultaneous explosive attacks in a highly populated area. Members of VVV have carried out attacks of this nature before.

In the last several months, both local authorities and the FBI have increased surveillance of professional basketball, baseball, football, and hockey events in the Los Angeles area. On one occasion, a suspicious package was left in a crowded area at a professional basketball game. The package turned out to be a hoax, but several authorities reported suspicious persons possibly observing the response. There is no way to be sure, but the hoax package could have been used to test the response of security and law enforcement.

Last week, the FBI confiscated financial statements and froze the bank accounts of a Los Angeles lawyer suspected of partially supporting members of VVV in the US. In the financial statements were records of a recent purchase of "military materials". It is unknown what exactly was purchased or where the materials are located now.

Summary

The militant group VVV might use explosives to attack a professional sporting event in Los Angeles. If the attack occurs, a plausible worst-case scenario would be 1000 deaths and injuries and 50 million dollars in property damage.

Based on the evidence outlined above and our professional judgment and experience, we estimate that the probability that this attack will occur over the next six months is 10%

Confidence condition: [...we are 5% sure that this attack will occur over the next six months.]

Range condition: [... our best estimate of the probability that this attack will occur over the next six months is 10%, but the probability could be as low as 1% or as high as 20%.]

4. How valuable is this intelligence report? In other words, if you had to decide what should be done about this attack, how useful or valuable is this report for determining future actions to take?

| Not at all valuable | | | | | | Fairly valuable | | | | | | Extremely valuable |
|---------------------------|---|---|---|---|---|--------------------|---|---|---|----|--|-----------------------|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |

5. Below are several questions about how you feel about the information and conclusions presented by the analysts. Please circle the number between the pair of words that best describes how you feel about the information and conclusions presented in the intelligence report.

| | | | | | | | | | | | | |
|-------------------------------------|---|---|---|---|---|---|---|---|---|---|----|--------------------------|
| Can't be trusted | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Can be trusted |
| Is inaccurate | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Is accurate |
| Is unfair | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Is fair |
| Doesn't tell whole story | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Tells whole story |
| Is biased | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Is unbiased |

FURTHER QUESTIONS ABOUT THE INTELLIGENCE REPORTS
[These questions were asked at the end of the experiment.]

1. Now focus specifically on the evidence that is presented in each intelligence report. How credible is the evidence overall? By “credible” we mean the ability to trust or believe the evidence. For example, people often feel that something they have “seen with their own two eyes” is more credible than a rumor they heard from a stranger.

How would you rate the overall credibility of the evidence presented in each report? Make a separate rating for each of the four intelligence reports. Feel free to go back and look at the reports again.

| | Very little credibility | | | | Moderate credibility | | | | Very high credibility | | | |
|------------------|----------------------------|---|---|---|-------------------------|---|---|---|--------------------------|---|----|--|
| Report #1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #4 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| | Very little credibility | | | | Moderate credibility | | | | Very high credibility | | | |

2. Again, focus specifically on the evidence that is presented in each intelligence report. How well does the evidence fit into a coherent story? By a “coherent story” we mean the ease to which you can form a good story or scenario from the evidence. For example, if all of the evidence fits into a believable story and there are not any other plausible explanations for the evidence, then you would rate the evidence as being very coherent. If, on the other hand, some pieces of evidence fit into a story but others do not, or there is more than one plausible story that fits the evidence, then you would make a lower rating for the coherence of the evidence.

How would you rate the overall coherence of the evidence presented in each report? Make a separate rating for each of the four intelligence reports. Feel free to go back and look at the reports again.

| | Very little coherence | | | | Moderate coherence | | | | Very high coherence | | | |
|------------------|-----------------------|----------|----------|----------|--------------------|----------|----------|----------|---------------------|----------|-----------|--|
| Report #1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #4 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| | Very little coherence | | | | Moderate coherence | | | | Very high coherence | | | |

APPENDIX D

MATERIALS FOR STUDY 4

Note: The experimental design and the terrorist scenarios were the same as those used in Study 3. In Study 4, however, there was an additional brief summary detailing how each terrorist attack had occurred several weeks earlier. This brief summary preceded each intelligence forecast. Below are each of these summaries. The dependent variables for Study 4 were nearly identical to those used in Study 3, however the wording is slightly changed because the judgments are taking place in hindsight.

TERRORIST ATTACK: SCENARIO #1

Instructions: Please read the following paragraph about a terrorist attack that occurred several weeks ago.

Summary of the attack:

Several weeks ago, a bomb was detonated on a passenger train in Chicago killing over 900 people and wounding hundreds more. It has become clear that militant group YYY was responsible for the attack. The attack would most likely have been stopped if additional security had been assigned to protect targets in Chicago. A special congressional committee has been formed and several politicians have begun criticizing the intelligence community.

Turn to the next page to read an intelligence report that was submitted to senior decision makers three weeks before this attack. You will then be asked to make a series of judgments about this intelligence report.

[Participants then read an intelligence forecast that was submitted to decision makers before this attack occurred]

TERRORIST ATTACK: SCENARIO #2

Instructions: Please read the following paragraph about a terrorist attack that occurred several weeks ago.

Summary of the attack:

Several weeks ago, a bomb was detonated outside of a federal building in Washington, DC killing over 900 people and wounding hundreds more. It has become clear that militant group XXX was responsible for the attack. The attack would most likely have been stopped if additional security had been assigned to protect targets in Washington, DC. A special congressional committee has been formed and several politicians have begun criticizing the intelligence community.

Turn to the next page to read an intelligence report that was submitted to senior decision makers three weeks before this attack. You will then be asked to make a series of judgments about this intelligence report.

[Participants then read an intelligence forecast that was submitted to decision makers before this attack occurred]

TERRORIST ATTACK: SCENARIO #3

Instructions: Please read the following paragraph about a terrorist attack that occurred several weeks ago.

Summary of the attack:

Several weeks ago, a bomb was detonated in a crowd at a sporting event in Los Angeles killing over 900 people and wounding hundreds more. It has become clear that militant group VVV was responsible for the attack. The attack would most likely have been stopped if additional security had been assigned to protect targets in Los Angeles. A special congressional committee has been formed and several politicians have begun criticizing the intelligence community.

Turn to the next page to read an intelligence report that was submitted to senior decision makers three weeks before this attack. You will then be asked to make a series of judgments about this intelligence report.

[Participants then read an intelligence forecast that was submitted to decision makers before this attack occurred]

TERRORIST ATTACK: SCENARIO #4

Instructions: Please read the following paragraph about a terrorist attack that occurred several weeks ago.

Summary of the attack:

Several weeks ago, a bomb was detonated on a passenger ship in New York City killing over 900 people and wounding hundreds more. It has become clear that militant group ZZZ was responsible for the attack. The attack would most likely have been stopped if additional security had been assigned to protect targets in New York City. A special congressional committee has been formed and several politicians have begun criticizing the intelligence community.

Turn to the next page to read an intelligence report that was submitted to senior decision makers three weeks before this attack. You will then be asked to make a series of judgments about this intelligence report.

[Participants then read an intelligence forecast that was submitted to decision makers before this attack occurred]

Questions about the Intelligence Report
[These questions were asked after each scenario]

INSTRUCTIONS: Imagine that you had read the intelligence report 3 weeks before the eventual attack. Please answer the following questions about this intelligence report on the previous page. Feel free to look back at the intelligence report when making your ratings.

1. Judging from the intelligence report, what would have been your impression of the risk associated with this possible attack?

| | | | | | | | | | | |
|------------------|---|---|---|---|------------------|---|---|---|---|-------------------|
| Very low risk | | | | | Moderate risk | | | | | Very high risk |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

2. Judging from the intelligence report, what would have been your impression of the chance that this attack would occur over the next 6 months?

| | | | | | | | | | | | | | | | | | | | | |
|-----------|---|----|----|----|----|----|----|--------------------------|----|----|----|----|----|----|----|----|----|----|----|---------|
| No chance | | | | | | | | As likely as unlikely | | | | | | | | | | | | Certain |
| 0% | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 | 100% |

3. Focus on the potential outcome of the terrorist attack described in the intelligence report. Judging from the intelligence report, what is your impression of the overall harm that would be inflicted on people, property, the economy, etc, if the attack occurred?

| | | | | | | | | | | |
|--------------------------|---|---|---|---|-----------------------|---|---|---|---|----------------------|
| Not harmful at all | | | | | Moderately harmful | | | | | Extremely harmful |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

4. How valuable was this intelligence report? In other words, if you had to decide what should have been done about this possible attack, how useful or valuable would this report have been to you?

| | | | | | | | | | | | | |
|---------------------------|---|---|---|---|---|--------------------|---|---|---|----|--|-----------------------|
| Not at all valuable | | | | | | Fairly valuable | | | | | | Extremely valuable |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |

5. Below are several questions about how you feel about the information and conclusions presented in the intelligence report. Please circle the number between the pair of words that best describes how you feel about the information and conclusions presented in the intelligence report.

| | | | | | | | | | | | | |
|-------------------------------------|---|---|---|---|---|---|---|---|---|---|----|--------------------------|
| Can't be trusted | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Can be trusted |
| Is inaccurate | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Is accurate |
| Is unfair | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Is fair |
| Doesn't tell whole story | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Tells whole story |
| Is biased | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Is unbiased |

6. Think about both the intelligence report and the terrorist attack that occurred three weeks later. Some people are blaming the intelligence community for not doing a good job predicting whether this attack would occur. How much blame do you think should be placed on the analysts that produced the intelligence report?

| | | | | | | | | | | | |
|----------------------|---|---|---|---|-------------------|---|---|---|---|----|--------------------------------|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Very little Blame | | | | | Moderate Blame | | | | | | Great amount of Blame |

FURTHER QUESTIONS ABOUT THE INTELLIGENCE REPORTS
[These questions were asked at the end of the experiment]

1. Now focus specifically on the evidence that is presented in each intelligence report. How credible is the evidence overall? By “credible” we mean the ability to trust or believe the evidence. For example, people often feel that something they have “seen with their own two eyes” is more credible than a rumor they heard from a stranger.

How would you rate the overall credibility of the evidence presented in each report? Make a separate rating for each of the four intelligence reports. Feel free to go back and look at the reports again.

| | Very little credibility | | | | Moderate credibility | | | | Very high credibility | | | |
|------------------|----------------------------|----------|----------|----------|-------------------------|----------|----------|----------|--------------------------|----------|-----------|--|
| Report #1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Report #4 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| | Very little credibility | | | | Moderate credibility | | | | Very high credibility | | | |

2. Again, focus specifically on the evidence that is presented in each intelligence report. How well does the evidence fit into a coherent story? By a “coherent story” we mean the ease to which you can form a good story or scenario from the evidence. For example, if all of the evidence fits into a believable story and there are not any other plausible explanations for the evidence, then you would rate the evidence as being very coherent. If, on the other hand, some pieces of evidence fit into a story but others do not, or there is more than one plausible story that fits the evidence, then you would make a lower rating for the coherence of the evidence.

How would you rate the overall coherence of the evidence presented in each report? Make a separate rating for each of the four intelligence reports. Feel free to go back and look at the reports again.

| | Very little coherence | | | Moderate coherence | | | | Very high coherence | | | |
|------------------|-----------------------|----------|----------|--------------------|----------|----------|----------|---------------------|----------|----------|-----------|
| Report #1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Report #2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Report #3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Report #4 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | Very little coherence | | | Moderate coherence | | | | Very high coherence | | | |

APPENDIX E

DETAILS OF STATISTICAL ANALYSES

Study 2

Assumptions of the Univariate General Linear Model

A univariate general linear modeling framework was used to test the hypothesized effects on perceived risk. There were no missing data on the variables of interest. In addition, several of the assumptions of the univariate general linear model were assured by proper sampling practices and the roughly equal sample sizes within groups. These assumptions include independence and an identical within group error distribution. The univariate general linear model is robust against violations of the homogeneity of variance assumption with relatively large sample sizes and roughly equal sample sizes among the groups (i.e. less than 2:1). Finally, residual plots for each model were examined for nonlinearities and other indicators of poor model fit, as well as confirmation of a roughly normal distribution of errors and equal variance of errors across levels of the independent variables. No concerning violations of the assumptions were found.

Effect Size, Power, and Confidence Intervals

In general, I have tried to focus on effect sizes and the precision of estimation (i.e. reporting confidence intervals), as opposed to null hypothesis significance testing. However, p-values are reported for the bulk of the statistical results. In addition, r is used as an effect size measure in Study 2. There are several other alternatives, for example Cohen's d , but r has several advantages over standardized difference measures of effect size. The primary benefit of r is the generality of interpretation as a measure of the linear relationship between two variables (Rosenthal, 1994). For example, r makes conceptual sense whether the variables of interest are both continuous in nature (Pearson's r), or one is dichotomous and one is continuous (Point-biserial). Mean difference indexes (e.g.

Cohen's d) make sense in the later case, but are not intuitively meaningful when both variables are continuous. Although the interpretation of any effect size must be understood within the context of the particular application, Cohen (1988) has developed general guidelines for interpreting r effect sizes: Small: $r=.1$, Medium: $r=.3$, Large: $r=.5$.

The concept of statistical power is a very important, and often ignored, aspect of statistical analysis. Rough power analyses were conducted to assure sufficient power for the primary effects of interest during the design of this experiment. Like any study, however, additional a priori hypotheses are often developed after the study design is finalized, and additional post-hoc research questions are often of interest once the analysis stage begins. In these cases, precise post-hoc power estimates are often difficult to compute. Thus, 95% confidence intervals are included to give the reader a general idea of the precision of estimation (i.e. statistical power) in the parameters of interest (Loftus, 2004). In general, the smaller the confidence intervals the greater the precision and the higher the statistical power.

Multivariate GLM Assumptions

As in the univariate case, a large sample size and roughly equal cell sizes ensure robustness of the multivariate solution against the violation of the multivariate normality and homogeneity of variance-covariance matrices assumptions. Scatterplots were used to assess the linearity assumption – namely, that all of the dependent variables are linearly related. Examination of the scatterplots revealed no nonlinear relationships of concern.

Multivariate Effect Size and Confidence Intervals

There are several different ways to represent the magnitude of individual model effects within the context of the multivariate general linear model (Kline, 2004). Pillai's V or Wilks' lambda are common choices and represent the proportion of explained and unexplained variance, respectively. In an effort to keep the effect size measures comparable across the univariate and multivariate analyses in Study 1, r effect sizes are reported for the multivariate effects. The r effect size can be computed in several

different ways – namely, by taking the square root of the Pillai's V statistic for the effect of interest, or by calculating a discriminant function score for each participant and then calculating the r effect size from these scores in the same manner as in the univariate case. For the majority of the effects of interest the two methods of computation converged, but in some cases there were discrepancies (e.g. when conducting simple effect tests). In those cases the r effect size calculated from the discriminant function scores is reported because I feel it is a more accurate representation of the effect size for specific simple effects. The Pillai's V in the simple effect procedure in SPSS is controlling for all other comparisons in the model, and consequently, produces a slightly different Pillai's V that when raised to the power of 1/2 is not a good representation of the effect size for the contrast of interest.

Study 3 & Study 4

Multilevel Models

Repeated measures designs are not optimally modeled with the General Linear Modeling (GLM) framework that was used to analyze the fully between subjects data in Study 2. A further generalization of the GLM called a Linear Mixed Model is more appropriate for data structures with repeated measurements. The subspecies of linear mixed models are known as multilevel mixed models, hierarchical linear models (HLM), or random-effects models.

There are several reasons why a multilevel framework is considered superior to a GLM for repeated measures data: 1) Multilevel models are more flexible in terms of data requirements (e.g. the repeated measures do not need to be measured at the same time for each subject), 2) multilevel models permit more control over the covariance structure, and 3) it is easier to work with time-varying covariates in the context of multilevel models (Raudenbush & Bryk, 2002). For these reasons, a multilevel framework was used

to model the effects of the experimental manipulations, as well as the effects of the subject-level and time-varying covariates.

General Specification of Multilevel Models and Model Building

Two-level models

All of the multilevel models used in Studies 3 and 4 were two-level models with the repeated measures data modeled at level 1 and the between subjects data model at level 2. For example, the following model was fit to assess the impact of stated likelihood, properties of the narrative evidence summary, and the moderating influence of stated likelihood format and consumer numeracy on perceptions of likelihood in Study 3 (the results of this model are presented in Table 5.10):

| | |
|---|---|
| LEVEL 1 MODEL | (bold: group-mean centering; bold italic: grand-mean centering) |
| $CHANCE_L = \beta_0 + \beta_1(LINEAR_T) + \beta_2(EVI_PROP) + r$ | |
| LEVEL 2 MODEL | (bold italic: grand-mean centering) |
| $\beta_0 = \gamma_{00} + \gamma_{01}(H1_COND) + \gamma_{02}(H2_COND) + \gamma_{03}(NUMTOT) + u_0$ | |
| $\beta_1 = \gamma_{10} + \gamma_{11}(H1_COND) + \gamma_{12}(H2_COND) + \gamma_{13}(NUMTOT)$ | |
| $\beta_2 = \gamma_{20} + \gamma_{21}(H1_COND) + \gamma_{22}(H2_COND) + \gamma_{23}(NUMTOT)$ | |

At level 1, the dependent variable is the likelihood ratings from each consumer for each of the three intelligence forecasts that they read. Therefore, each consumer expressed his or her perceived likelihood to a forecast with a stated likelihood of 1%, a forecast with a stated likelihood of 5%, and a forecast with a stated likelihood of 10% (the pure narrative forecast is not included in this analysis). These perceptions of likelihood are then modeled as a function of a linear trend across the levels of stated likelihood (Linear_T in the figure above). Consumers also rated the coherence and credibility of each of the three forecasts that they read. Thus, perceptions of likelihood are also modeled as a function of consumer's ratings of the credibility and coherence of each forecast (Evi_Prop in the figure above). Three parameters are then estimated from

this level 1 model: β_0 = the intercept, or the mean level of perceived likelihood averaging across stated likelihood and at the mean level of credibility and coherence; β_1 =the extent to which likelihood perceptions show a linear trend across stated likelihood at the mean level of credibility and coherence; β_2 =the linear relationship between perceptions of credibility and coherence and perceptions of likelihood averaging across stated likelihood.

At level 2, these three parameters become dependent variables and variance in these parameters for each consumer are modeled as a function of the between subject variables. In this case, the between-subject variables are the format of the stated likelihood in the forecast, which are represented as two helmert contrasts (H1=contrast between the range condition and the two point estimate conditions; H2=contrast between the point estimate conditions, internal and external framing of likelihood), and the total score on the numeracy individual difference measure. The goal of the level 2 model is to see if the variance in the parameters estimated at level 1 can be predicted by the between subject variables represented at level 2. For instance, parameter γ_{13} at level 2 is an estimate of the extent to which the effect of stated likelihood on perceived likelihood (averaging across likelihood format) can be predicted by the numeracy level of the consumer. In other words, this is a test of a cross-level interaction.

Model building

In general, all of the multi-level models used in studies 3 and 4 were specified to test specific hypotheses of interest. However, exploratory analyses were also conducted to test for higher-order interaction effects that would clarify the effects found elsewhere in the models. Higher order interactions that were not significant were removed and the more parsimonious model results are reported.

Assumptions of Multilevel Models and Standard Error

Each of the models reported above were fit with Restricted Maximum Likelihood Estimation (REML). Similar statistical assumptions underlie parameter estimation in multilevel models and multiple regression analysis, although in the case of multilevel

models there is multilevel data structure. Violations of critical assumptions can negatively influence standard errors and inferential tests (Raudenbush & Bryk, 2002). Residual plots for each model were examined for nonlinearities, outliers, and other indicators of poor model fit, as well as confirmation of a roughly normal distributions of errors and equal variance at each level of the multilevel model. No concerning violations of the assumptions were found in any of the models fit above. Furthermore, all inferential tests are conducted with robust standard errors, which further guard against the influence of violating critical assumptions (Raudenbush & Bryk, 2002, pg. 276).

Random versus Fixed Coefficients

In a two level model, predictors at level-1 and level-2 are modeled as fixed effects. However, the intercepts and slopes that are estimated at level-1 can be modeled as fixed, non-randomly varying, or randomly varying (Raudenbush, Bryk, Cheong, Congdon & du Toit, 2004). A fixed intercept or slope means that the parameter is assumed to be equal for each level two unit, which in this case is the individual consumers. A non-randomly varying intercept or slope means that the parameter is expected to vary across level-2 units with respect to level-2 predictor variables, but does not vary randomly for each individual. For example, in Study 2 one of the primary predictions was that the linear effect across probability levels on perceived likelihood would be moderated by the consumer numeracy (a level-2 variable). Randomly varying intercepts or slopes means that these parameters are a function of overall population effects as well as a “random”, or unique contribution for each person. Of course, an intercept or a slope can also be modeled as having contributions from unique, non-random sources (e.g. numeracy) as well as unique, random effects for each person.

In each of the multilevel model estimates above, the intercept is modeled with both random and non-randomly varying components. The intercepts were modeled as random because I wanted to capture the idiosyncratic (random) way in which consumers may be using the ratings scales. For example, some consumers may show similar slopes across the within subject factor in terms of perceived likelihood, but they just start at

different places on the rating scale. This might happen if, for instance, the likelihood scale was interpreted as a more general rating scale to which consumers scaled their responses in an idiosyncratic way. However, the slope terms were modeled as only non-randomly varying (i.e. as functions of the level-2 variables), without random error terms. This was done because I had no a priori reason to assume random variation in the slopes, and practically, with only 3-4 levels of the within subject factor in a given model it was often impossible to estimate all of the random effects (i.e. there were not enough degrees of freedom).

Not allowing the slope coefficients to vary randomly introduces a potential misspecification problem, if in fact the slope coefficients do have substantial random variance components. One way to assess the impact of the potential misspecification of random effects is to compare model-based and robust estimates of standard error in the model without the random variance component (Raudenbush & Bryk, 2002). If the standard error estimates are substantially different from one another, then this is an indication that the fixed coefficients may need to be specified as random. The model-based and robust standard errors were not substantially different in any of the models reported above. In addition, the specification of slope coefficients as random (in models with adequate degrees of freedom) did not result in any substantive differences in the interpretation of effects from a non-randomly varying specification.

Centering

It is very important in multilevel modeling that each of the level-1 and level-2 predictors are represented in a way that makes the coefficients scientifically interpretable. As in standard regression analyses, this is achieved through centering, or specifying the location for, the level-1 and level-2 variables (Raudenbush & Bryk, 2002). There are several different ways to center predictor variables, with grand mean centering or group mean centering as the most common options. In the models reported above, each of the continuous and categorical variables (often represented by contrast coded dummy variables) are grand mean centered. This results in a similar interpretation as the standard

ANCOVA model, for instance, where the intercept in the level-1 model is interpreted as the grand mean adjusted for any covariates in the model (Raudenbush & Bryk, 2002). For example, grand mean centering is useful in specific cases where “main effect” type analyses are of interest, like a specific contrast on the probability format variable averaging across level of the other experimental factor (probability level).

Effect Size

Ideally, every effect estimated in the multilevel models could be represented by a common effect size measure, like Cohen’s d or r . However, since multilevel modeling is a relatively new statistical approach, methodologists are still actively searching for the most appropriate ways to represent effect sizes for individual model parameters. Since r effect sizes were reported in Study 2, I tried various methods of calculating r effect sizes for the parameters in the multilevel analyses to make them comparable with the effects from Study 2. Since there is relative dearth of literature on the calculation of the effect sizes in these models, I was not able find a consistent way to calculate r effect sizes that didn’t leave me with the lingering feeling that I was doing something wrong. In the end, I decided to follow the lead of Raudenbush & Xiao-Feng (2001) and Tymms, Merrell, and Henderson (1997), who present similar approaches for calculating effect sizes in multilevel models. The basic approach is to generalize the standardized difference effect size measures discussed by Cohen (1988) and Glass (1981) to the multilevel context. In short, for dummy coded categorical variables and standardized continuous variables the effect size takes the following general form:

$$\Delta_p = \frac{\beta_{p1}}{\sqrt{\tau_{pp}}},$$

where Δ_p is the standardized effect size measure, β_{p1} is a specific coefficient for the effect of interest, and $\sqrt{\tau_{pp}}$ is the square root of an appropriate random variance component. For example, assume that the primary effect of interest at level-1 is the linear slope between the narrative and numerical probability level conditions in Study 3.

Assume that β_{p1} is the coefficient indicating the difference between two probability format conditions (between subject variable at level-2) on this linear slope, in other words it is a cross level interaction effect. The effect size Δ_p , therefore, is a representation of the difference between the two conditions on this linear slope, β_{p1} , divided by the population variation in the linear slopes (random variance component), and can be considered a standardized mean difference.

Since these effect size measures have not been thoroughly studied, they should not be directly compared to other standardized mean difference effect sizes (e.g. Cohen's d), or the r effect sizes calculated in Study 2 that could readily be converted into Cohen's d . The main reason for presenting the effect size measures in Study 2 and Study 3 is to provide a framework for comparing the magnitude of the different effects within each study. So, for instance, the reader can compare the magnitude of important hypothesized effects like the effect of stated probability information compared to the effect of the properties of the narrative evidence on perceived likelihood.

Power Analysis

Under the assumption that the GLM was going to be used for analysis, a rough power analysis was conducted to estimate the sample size needed to test the between and within subject effects with adequate statistical power. Assuming between a small and medium effect size (using Cohen's criteria), a total sample size of $N=60$ would produce power = .80 for the between subjects comparisons of interest. For the within subject effect, a total sample size of $N=60$ would produce power = .96. However, this sample size may have been too small to reliably detect individual differences associated with the numeracy. Correlations between numeracy and the dependent variables in Study 1 ranged from $r = .20-.25$. Assuming a correlation of $r = .25$ and a sample size of $N=60$, estimated statistical power = .65. If the total sample size was increased to $N=80$, then power=.80 for detecting a simple correlation with numeracy. At least from the perspective of the GLM, the sample sizes of $N=87$ and $N=81$ appear sufficient.

However, formal power analysis was not conducted within the context of multilevel models. For multilevel models, in general, the number of groups has more effect on statistical power than the number of observations. In this case the individual participants are the group level variable, and it is recommended that the higher-level sample size (level-2) is at least 20, but preferably 50 for adequate statistical power (Garson, 2007). Thus, I expect to have adequate statistical power with $N=87$ and $N=81$ at level-2.

REFERENCES

- Armstrong, J. S. (Ed.) (2001). *Principles of Forecasting*. Boston: Kluwer Academic Publishers.
- Armstrong, W. C., Leonhart, W., McCaffery, W. J., & Rothenberg, H. C. (1995). The hazards of single-outcome forecasting. In H.B. Westerfield (Ed.) *Inside CIA's private world* (pp. 238-254). New Haven: Yale University Press.
- Beach, L. R., & Braun, G. P. (1994). Laboratory studies of subjective probability: A status report. In G. Wright & P. Ayton (Eds.) *Subjective Probability* (pp. 107-127). Chichester: John Wiley & Sons.
- Bisantz, A. M., Marsiglio, S. S., & Munch, J. (2005). Displaying uncertainty: Investigating the effects of display format and specificity. *Human Factors*, 47(4), 777-796.
- Brun, W. (1994). Risk perception: Main issues, approaches and findings. In G. Wright & P. Ayton (Eds.) *Subjective Probability* (pp. 185-209). Chichester: John Wiley & Sons.
- Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, D. L. Medin, & R. Hastie (Eds.), *Decision making from a cognitive perspective*. New York: Academic.
- Burkell, J. (2004). What are the chances? Evaluating risk and benefit information in consumer health materials. *Journal of the Medical Library association*, 92(2), 200-208.
- Clark, R. M. (2004). *Intelligence analysis: A target centric approach*. Washington, DC: CQ Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates.
- Cooper, J. R. (2005). *Curing analytic pathologies: Pathways to improves intelligence analysis*. Monograph: Center for the Study of Intelligence.

- Curley, S. P., & Benson, P. G. (1994). Applying a cognitive perspective to probability construction. In G. Wright & P. Ayton (Eds.) *Subjective Probability* (pp. 185-209). Chichester: John Wiley & Sons.
- Dieckmann, N. (2007). Numeracy: A review of the literature. Report submitted to the National Cancer Institute. (available from author on request).
- Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, *66*, 819-829.
- Epstein, R. M., Alper, B. S., & Quill, T. E. (2004). Communicating evidence for participatory decision making. *Journal of the American Medical Association*, *291*(19), 2359-2366.
- Fagerlin, A., Ubel, P. A., Smith, D. M. & Zikmund-Fisher, B. J. (submitted). Making numbers matter: Present and future research in risk communication.
- Fischhoff, B. (1975). Hindsight \neq Foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288-299.
- Fischhoff, B. (2001). Learning from experience: Coping with hindsight bias and ambiguity. In J.S. Armstrong (Ed.) *Principles of forecasting: A handbook for researchers and practitioners* (pp. 543-554). Boston: Kluwer Academic Publishers.
- Fisk, C. E (1995). The sino-soviet border dispute: A comparison of the conventional and Bayesian methods for intelligence warning. In H.B. Westerfield (Ed.) *Inside CIA's private world* (pp. 264-273). New Haven: Yale University Press.
- Flugstad, A. R., & Windschitl, P. D. (2003). The influence of reasons on interpretations of probability forecasts. *Journal of Behavioral Decision Making*, *16*, 107-126.
- Fox, C. R., & Malle, B. F. (1997). On the communication of uncertainty: Two modes of linguistic expression. Unpublished manuscript.
- Fox, C. R., & Irwin, J. R. (1998). The role of context in the communication of uncertain beliefs. *Basic and Applied Social Psychology*, *20*(1), 57-70.
- Garrick, B. J. (2002). Perspectives on the use of risk assessment to address terrorism. *Risk Analysis*, *22*(3), 421-423.

- Garson, D. G. (2007). *Linear mixed models*. Retrieved August 28, 2007 from <http://www2.chass.ncsu.edu/garson/pa765/multilevel.htm>
- Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.) *Subjective Probability* (pp. 129-161). Chichester: John Wiley & Sons.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, *103*(3), 592-596.
- Glass, G.V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. London: Sage.
- Gregory, W. L., & Duran, A. (2001). Scenarios and acceptance of forecasts. In J.S. Armstrong (Ed.) *Principles of forecasting: A handbook for researchers and practitioners* (pp. 517-540). Boston: Kluwer Academic Publishers.
- Gurmankin, A. D., Baron, J., & Armstrong, K. (2004). The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. *Medical Decision Making*, *24*(3), 265-271.
- Haimes, Y. Y. & Longstaff, T. (2002). The role of risk analysis in the protection of critical infrastructures against terrorism. *Risk Analysis*, *22*(3), 439-444.
- Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks: Sage Publications.
- Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgment of past events after outcomes are known. *Psychological Bulletin*, *107*(3), 311-327.
- Helton, J. C. (1993). Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliability Engineering and System Safety*, *42*, 327-367.
- Hendrickx, L., Vlek, C., & Calje, H. (1992). Effects of frequency and scenario information on the evaluation of large-scale risks. *Organizational Behavior and Human Decision Processes*, *52*, 256-275.
- Hendrickx, L., Vlek, C., & Oppenwal, H. (1989). Relative importance of scenario information and frequency information in the judgment of risk. *Acta Psychologica*, *72*, 41-63.

- Heuer, R.J. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence Analysis: Central Intelligence Agency.
- Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290, 2261-2262.
- Horowitz, B. M., & Haimes, Y. Y. (2003). Risk-based methodology for scenario tracking, intelligence gathering, and analysis for countering terrorism. *Systems Engineering*, 6(3), 152-169.
- Hsee, C.K. (1995). Elastic justification: How tempting but task-irrelevant factors influence decisions. *Organizational Behavior & Human Decision Processes*, 62(3), 330-337.
- Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67, 247-257.
- Johnson, B. B. (2003). Further notes on public response to uncertainty in risks and science. *Risk Analysis*, 23(4), 781-789.
- Johnson, B. B., & Slovic, P. (1995). Presenting uncertainty in health risk assessment: Initial studies of its effects on risk perceptions and trust. *Risk Analysis*, 15(4), 485-494.
- Johnson, B. B., & Slovic, P. (1998). Lay views on uncertainty in environmental health risk assessment. *Journal of Risk Research*, 15, 261-279.
- Kahan, D. M., Braman, D., Slovic, P., Gastil, J., & Cohen, G. (2007). The second national risk and culture study: Making sense of, and making progress in, the American culture war of fact. Report by The Cultural Cognition Project at Yale Law School. <http://research.yale.edu/culturalcognition/>
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1982a). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 509-520). Cambridge: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1982b). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 201-208). Cambridge: Cambridge University Press.

- Keller, C., Siegrist, M., & Gutscher, H. (2006). The role of the affect and availability heuristics in risk communication. *Risk Analysis*, 26(3), 631-639.
- Kline, R. B. (2004). *Supplemental chapter on multivariate effect size estimation*. Retrieved June 5, 2006 from <http://www.apa.org/books/resources/kline>
- Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21, 37-44.
- Loftus, G. R. (2004). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler, S. Yantis, D. Medin, R. Gallistel & J. Wixted (Eds.) *Stevens' handbook of experimental psychology*, 3rd Edition (Chapter 9). New Jersey: John Wiley & Sons.
- McComas, K. A., & Trumbo, C. W. (2001). Source credibility in environmental health-risk controversies: Application of meyer's credibility index. *Risk Analysis*, 21(3), 467-480.
- Morgan, M.G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.
- Pate-Cornell, M. E. (1996). Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering and System Safety*, 54, 95-111.
- Pate-Cornell, M. E. (2002). Fusion of intelligence information: A bayesian approach. *Risk Analysis*, 22(3), 445-454.
- Paulos, J. A. (Ed.). (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.
- Pennington, N., & R. Hastie. 1988. Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 14: 521-533.
- Pennington, N., & Hastie, R. (1994). A theory of explanation-based decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsombok (Eds.) *Decision making in action: Models and methods* (pp. 188-201). New Jersey: Ablex Publishing Corporation.
- Peters, R. G., Coviello, V. T., & McCallum, D. B. (1996). The determinants of trust and credibility in environmental risk communication: An empirical study. *Risk Analysis*, 17(1), 43-54.

- Peters, E. (in press). Preferred data visualization techniques may not lead to comprehension and use of hazard information: Commentary on Pang. In A. Bostrom, S. P. French, & S. J. Gottlieb, (Eds.), *Risk Assessment, modeling and decision support: Strategic directions*. Heidelberg, Germany: Springer.
- Peters, E., Vastfjäll, D., Slovic, P., Mertz, C., Mazzocco, K., & Dickert, S. (2006). Numeracy and Decision Making. *Psychological Science, 17*(5), 407-413.
- Peters, E, Dieckmann, N.F., Dixon, A., Slovic, P., Mertz, C.K., Slovic, P., & Hibbard, J. H (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review, 64*(2), 169-190.
- Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C.K., Slovic, P., & Hibbard, J. H. (2006). Bringing meaning to numbers: The effects of affect in choice. Manuscript submitted for publication.
- Politi, M. C., Han, P.K.J., & Col, N. (2006). Communicating the uncertainty of harms and benefits of medical interventions. *Eisenberg center 2006 white paper series*.
- Price, P. C.. & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17*, 39-57.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. K., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Raudenbush, S. W., & Xio-Feng, L. (2001). Effect of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. M. Cooper & L. V. Hedges (Eds.) *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
- Rowe, W. D. (1994). Understanding uncertainty. *Risk Analysis, 14*(5), 743-750.
- Schrage, M. (2005). *What percentage is "Slam Dunk"?* Editorial in the Washington Post, February, 20th, 2005.

- Schwartz, L. M., Woloshin, S., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, 127(11), 966-972.
- Schwartz, P. (1996). *The art of the long view*. Bantam Doubleday Dell Publishing Group.
- Siegrist, M. (1997). Communicating low risk magnitudes: Incidence rates expressed as frequency versus rates expressed as probability. *Risk Analysis*, 17(4), 507-510.
- Slovic, P. (1987). Perception of risk. *Science*, 236, 280-285.
- Slovic, P., Monahan, J., & McGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, 24(3), 271-296.
- Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis*, 24, 311-322.
- Sunstein, C. R. (2003). Terrorism and probability neglect. *The Journal of Risk and Uncertainty*, 26(2/3), 121-136.
- Teigen, K. H. (1994). Variants of subjective probabilities: Concepts, norms, and biases. In G. Wright & P. Ayton (Eds.) *Subjective Probability* (pp. 211-238). Chichester: John Wiley & Sons.
- Tetlock, P. E. (2005). *Expert political judgment*. Princeton: Princeton University Press.
- Thompson, K. M. (2002). Variability and uncertainty meet risk management and risk communication. *Risk Analysis*, 22(3), 647-654.
- Trumbo, C. W., & McComas, K. A. (2003). The function of credibility in information processing for risk perception. *Risk Analysis*, 23(2), 343-353.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A., & Kahneman, D. (1982a). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 84-98). Cambridge: Cambridge University Press.

- Tversky, A., & Kahneman, D. (1982b). Causal schemas in judgments under uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 117-128). Cambridge: Cambridge University Press.
- Tymms, P., Merrell, C., & Henderson, B. (1997). The first year at school: A quantitative investigation of the attainment and progress of pupils. *Educational Research and Evaluation*, 3(2), 101-118.
- Vlek, C., & Stallen, P. J. (1981). Rational and personal aspects of risk. *Acta Psychologica*, 45, 273-300.
- Walker, V. R. (1995). Direct inference, probability, and a conceptual gulf in risk communication. *Risk Analysis*, 15(5), 603-609.
- Wallsten, T. S., & Bedescu, D. V. (1995). A review of human linguistic processing: General principles and empirical evidence. *The Knowledge Engineering Review*, 10, 43-62.
- Wallsten, T. S., Bedescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, 39, 176-190.
- Wasserman, D., Lempert, R. O. & Hastie, R. (1991). Hindsight and causality. *Personality and social psychology bulletin*, 17, 30-55.
- Windschitl, P. D., Martin, R., & Flugstad, A. R. (2002). Context and the interpretation of likelihood information: The role of intergroup comparisons on perceived vulnerability. *Journal of Personality and Social Psychology*, 82, 742-755.
- Windschitl, P. D., & Weber, E. U. (1999). The interpretation of “likely” depends on the context, but “70%” is 70% -- right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1514-1533.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Elsevier Academic Press.
- Yates, F. J., Price, P. C., Lee, J. & Ramirez, J. (1996). Good probabilistic forecasters: The ‘consumer’s’ perspective. *International Journal of Forecasting*, 12, 41-56.
- Zlotnick, J. (1995). Bayes’ theorem for intelligence analysis. In H.B. Westerfield (Ed.) *Inside CIA’s private world* (pp. 255-263). New Haven: Yale University Press.