# Communicating Weather and Climate Uncertainty:
# Exploratory Research in Cognitive Psychology

Jared LeClerc

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Susan L. Joslyn, Chair

John C. Palmer

John M. Miyamoto

Program Authorized to Offer Degree:

Psychology

University of Washington

## Abstract

Communicating Weather and Climate Uncertainty:
Exploratory Research in Cognitive Psychology

Jared LeClerc

Chair of the Supervisory Committee:
Susan L. Joslyn, PhD
Department of Psychology

Vast improvements in weather prediction, such as the ability to calculate precise and reliable estimates of weather uncertainty, would theoretically help members of the public make important decisions related to weather risks, but the extent to which non-expert end-users understand forecast uncertainty information remains unclear. A small body of recent research suggests that, contrary to the implications of much past research in cognitive psychology, non-experts can make effective use of probabilistic uncertainty estimates in weather forecasts. The research reported here uses an experimental framework to expand upon those findings to explore the scope of the benefits of inclusion of uncertainty estimates in forecasts. Several uncertainty expressions were tested in a variety of hypothetical settings using laboratory decision tasks based on real-world weather- and climate-related decision situations. Results consistently suggested that non-experts made better weather- and climate-related decisions when forecasts and projections included uncertainty estimates, although some expressions of uncertainty were more effective in some situations than in others. This research also explores effects of forecast error, false alarm effects, and trust. It adds to the growing body of research about how to effectively

communicate weather- and climate-related risk, with results that might generalize to other domains in which non-experts make difficult decisions when faced with uncertainty. The research adds to our understanding of basic psychology, as it explores such concepts as framing, perception of the likelihood of rare events, and the role of decision feedback.

**Table of Contents**

## Introduction

> *"What do we do with multiple models? Or if they're inconsistent? You might like having 61 models of a storm track, but [that's not helpful to us]. We need less data and more decision-level information."*
>
> – Major General Michael Walsh, U.S. Army Corps of Engineers, addressing atmospheric scientists at the 2013 Annual Meeting of the American Meteorological Society

In a wide variety of domains, from finance to sports to weather, there is concerted effort to create predictive models that incorporate increasingly staggering amounts of data (Silver, 2012a). Weather prediction, for instance, has improved dramatically over the past few decades (Silver, 2012b), largely due to new ways of modeling data (e.g., Gneiting & Raftery, 2005; Sloughter, Raftery, Gneiting, & Fraley, 2007) and increased computing power. More and more information of ever-higher quality is available for use. But can non-expert end-users make effective use of such information, much of which is probabilistic? If people were able to reason and make good decisions with this sort of probabilistic or otherwise complex information, then breakthroughs in forecasting science might directly and immediately benefit the general public: Better information, and more of it, could lead to better judgment and more informed decisions.

However, substantial empirical evidence from research in higher-order cognition suggests that humans are subject not only to reasoning errors, but to systematic patterns of such errors. Decades of research in cognitive psychology suggests that humans use mental shortcuts in judgment and decision making (Kahneman, 2011). These shortcuts, or heuristics, typically are adaptive but can also lead to errors (Tversky & Kahneman, 1974). While there is substantial ongoing debate among schools of thought in cognitive psychology about the implications of the weaknesses of heuristic reasoning (Klein, 2009; Gigerenzer, 2008), it has been demonstrated that

at least in certain circumstances, if only rarely, heuristic processes can lead people toward erred judgment. Importantly, this set of circumstances includes situations in which errors in judgment can have devastating consequences. One area that prominently features such situations is weather, the communication of weather-related risk, and weather-related decision making.

There are numerous weather-related risks. Certain events threaten entire communities, such as floods, hurricanes, and tornadoes, while other weather-related risks might be of specific concern to particular user groups, such as the risk of low precipitation to farmers, the risk of freezing temperatures to road maintenance managers, the risk of high wind to boaters, and even the risk of afternoon drizzle to party planners. While uncertainty is inherent in weather prediction, most forecast information that is communicated to the public includes only single-value estimates of weather parameters, suggesting that the forecasted weather events *will* occur. However, based on the substantial diversity of risk tolerances of people in different contexts (Joslyn & Savelli, 2010; Morss et al., 2008), it seems that including additional information about the likelihood of weather events occurring – that is, information that is inherent to those events – would benefit decision makers. Recent applied research suggests that people already understand that forecasts are uncertain, even if that uncertainty is not stated explicitly (Joslyn & Savelli, 2010; Morss et al., 2008), and that they make better weather-related decisions with uncertainty estimates than without them (Roulston et al., 2006; Nadav-Greenberg & Joslyn, 2009).

The body of research covered in this dissertation explores how best to communicate weather- and climate-related risk, with a particular focus on how well non-expert end-users understand probabilistic estimates of weather uncertainty. Recent research suggests that contrary to past assumptions, inclusion of probabilistic uncertainty estimates might enhance decision making, but the extent of that enhancement is not clear. Do decision makers benefit from

inclusion of uncertainty estimates in some situations but not in others? Is uncertainty information treated the same for far-future events (e.g., climate change) as for short-term weather events? Does explicit decision advice, which incorporates uncertainty estimates without making numeric probabilities known, help people make better decisions? These and other important real-world questions are explored in a novel experimental framework. This research features experimental evidence that expands upon recent findings concerning non-expert end-users' understanding of weather-related uncertainty.

This research contributes to two broad areas. First, this is applied research of critical real-world importance. I have taken real weather-related decision problems, many of which are discussed widely in the news media, and transformed them into controlled decision tasks, allowing me to systematically manipulate variables of interest. This type of rigorous, controlled research adds important information to field-based findings and ultimately could be used to improve everyday people's lives by way of informing better policies and engendering better decisions from emergency managers and others. And while the experiments reported here specifically concern weather- and climate-related judgment and decision making, many of the findings might well be generalizable to other areas in which non-experts are faced with difficult decisions involving uncertain future outcomes, such as medicine and finance. Second, this research addresses basic questions about human psychology, including questions about judgment and decision making and reasoning with probability. These are areas of great interest within the psychology research community and much debate exists about the mechanisms underlying these higher-order processes. In exploring the issues of weather- and climate-related judgment and decision making, the present research offers richer insight about psychological processes in general, including findings that likely extend well beyond the particular application of weather.

# Chapter 1. Background

## Goods Forecasts, Poor Decision Making

The science of weather prediction has improved dramatically in the past few decades. For example, in 1972, the National Weather Service's high temperature forecast three days out was off by 6°F on average; now it is off by only about 3°F (Silver, 2012b). From 1986 to 2004, tornado lead times dropped from an average of about five minutes to an average of 13 minutes, with a matching increase, from 25% to 75%, of tornados warned (Erickson & Brooks, 2006; Brotzge & Donner, 2013). In the mid-1980s, the National Hurricane Center would miss the point of a hurricane's landfall with three-days' notice by an average of about 350 miles; now it misses by an average of only about 100 miles (Silver, 2012b). These substantial improvements are due largely to developments in atmospheric modeling and increased computer power. Using Ensemble Prediction Systems (Eckel, Allen, & Sittel, 2012), forecast probabilities can now be calculated with an ever-increasing degree of accuracy and reliability (e.g., Gneiting & Raftery, 2005; Sloughter, Raftery, Gneiting, & Fraley, 2007).

Still, despite these improvements, substantial weather-related loss of life still occurs. For example, for the catastrophic EF-5 tornado that struck Joplin, MO, in May 2011, extremely timely and accurate warnings had been issued, and loss of human life could have been averted (Paul & Stimers, 2012). However, 158 people died in the event, making it the deadliest tornado in the United States in over 60 years (NOAA, 2011). In August 2005, despite the National Weather Service's almost exact prediction of Hurricane Katrina's landfall almost 2.5 days in advance (Silver, 2012b), there were more than 1,800 deaths, many of which could have been

avoided. And in October 2012, Superstorm Sandy directly caused the deaths of 147 people in the United States (NOAA, 2013), despite the extensive pre-storm coverage in the news media.

How people respond when threatened with a chance of extreme weather is an extremely complicated topic, and many factors affect people's decisions (e.g., Whitehead et al., 2000; Baker, 1991; Dash & Gladwin, 2007). However, in many of the extreme weather events in recent years that have resulted in tragedy, lives were not likely lost because of inaccurate forecasts, but rather because of poor decision making on the part of the people vulnerable to the weather events. There is ample evidence of people not taking sufficient precautionary action when faced with weather-related risk. For instance, in September 2008, as Hurricane Ike approached the Gulf Coast of the United States, fully 40% of residents did not comply with mandatory evacuation orders, despite being warned of "certain death" (McKinley & Urbina, 2008). Similarly, as Superstorm Sandy approached New York City in October 2012, approximately two-thirds of residents in certain highly vulnerable low-lying areas did not comply with evacuation orders (Gibbs & Holloway, 2013). It is not clear if the problem lies principally with the decision makers themselves or the way the risk was communicated to them. Either way, it is clear that the situation could be improved upon. Some of what underlies poor weather-related decision making can be explained by the contributions of research in cognitive psychology. Fortunately, psychology research also has the potential to identify ways to improve communication of weather uncertainty and to improve decision making, as well.

**Research on Judgment and Decision Making: Focus on Cognitive Limitations**

Much of the recent focus of research on the psychology of judgment and decision making has been on the limitations of human reasoning, notably on how people fail to make decisions

that meet the level of rational standards (see Hastie & Dawes, 2010). This has been mirrored in the popular press. Popular books like *Predictably Irrational* (Ariely, 2010), *The Invisible Gorilla: How Our Intuitions Deceive Us* (Chabris & Simons, 2011), and *Why We Make Mistakes* (Hallinan, 2010) have generally explored the disparity between "Econs" and "Humans" (Thaler & Sunstein, 2008). Econs make consistent, rational decisions consistent with classical economic theory, whereas Humans are inconsistent and illogical, often being influenced by numerous situational factors (Kahneman, 2011). The overall picture painted of humans' reasoning powers is negative and not encouraging (Klein, 2013).

This focus on cognitive limitations developed gradually over the past century. In the first half of the 20th century, the prevailing view was the rational agent model: People are essentially rational and are good intuitive statisticians, obeying elementary rules of probability, and errors in reasoning are typically unsystematic, computational errors (Gilovich & Griffin, 2002). For example, when making decisions, people assess the probabilities of possible outcomes, determine the utility (or personal value) associated with each outcome, combine these assessments, and choose the option with the highest combination of probability and utility (Gilovich & Griffin, 2002; von Neumann & Morgenstern, 1944). Decision theory axioms were regarded as not just normative but also were believed to describe the way most people actually made decisions (Kahneman, 2011). However, this belief began to erode when Paul Meehl (1954) performed a review of the accuracy of predictions made by clinical psychologists compared to statistical actuarial models and found that the algorithms outperformed the humans. A growing sense of the fallibility of human judgment was made even stronger by the research of psychologist Ward Edwards (1968), whose Bayesian analyses of judgment revealed that in everyday situations, people's judgments were not consistent with normative standards (Gilovich

& Griffin, 2002). Herb Simon (1957) suggested that the rational choice model was not a realistic standard for comparison and introduced the concept of bounded rationality, which posited that there were inherent processing limitations of the human mind. Bounded rationality suggested that "People reason and choose rationally, but only within the constraints imposed by their limited search and computational capacities" (Gilovich & Griffin, 2002, pg. 2) and that people used simplifying heuristics to compensate for cognitive limitations.

In the 1960s and 1970s, psychologists Daniel Kahneman and Amos Tversky developed a research program that demonstrated that processes of intuitive judgment are categorically different from the sort of normative reasoning processes formerly believed to govern judgment. They suggested that the processes of judgment were not just simpler than fully rational processes, but rather fundamentally different. They developed experiments that revealed biases in judgment, showing that people use heuristics, or mental shortcuts or rules of thumb, to make decisions. These heuristics made use of basic computations that the mind had evolved to make quickly and efficiently (Gilovich & Griffin, 2002). Kahneman and Tversky's research program, called the heuristics and biases approach, identified a number of heuristics that can bias judgments and guide decisions. While these heuristics are generally adaptive, they can potentially lead to systematic errors with severe consequences (Tversky & Kahneman, 1974). The basic approach of heuristics and biases research was to demonstrate a decision maker's failure to perform a task at the level of a normative model of rational choice, often by using an experimental task that in some way exploited the decision maker's reliance on a heuristic. This pattern of judgment, or bias, was then used to draw inferences about the cognitive mechanisms that underlie the judgment or decision-making process (Kahneman, 2011). The idea, then, was that understanding the shortcomings of a system often informed researchers of the way the

system operated. By developing experiments that revealed biases in judgment, Kahneman and Tversky were able to better understand the underlying heuristic processes and characterize the circumstances in which our reasoning is suboptimal (Kahneman, 2011). That errors in judgment could be systematic represented a fundamental shift from the preceding school of thought, in which errors made by rational actors were believed to be unsystematic (Gilovich & Griffin, 2002). And while many argue that heuristics are a generally efficient means to solve problems (Gigerenzer, 2008), the idea that most people are irrational has gained substantial traction, both in the academic literature and in the popular press.

Heuristic processes theoretically underlie decision making in many common real-life situations. An expansive literature, for example, explores decisions in the domain of medicine, focusing on the decision making of both patients (e.g., Lipkus, 2007) and providers (e.g., Elstein & Schwarz, 2002) and on how uncertainty is communicated and treatment options explained. A perhaps even more expansive literature explores financial decision making (see Gärling, Kirchler, Lewis, & van Raaij, 2009, for a recent review), again with interest in how professional investors and laypeople alike make decisions when faced with uncertainty. Researchers are also looking into weather- and climate-related decision making, examining how atmospheric scientists interpret weather uncertainty, how forecasters and emergency managers express it, and how non-expert forecast end-users interpret it. The present research focuses on the latter two groups. How should weather uncertainty be communicated? Do regular people understand weather uncertainty and weather- and climate-related risk? How do they use forecast information to make decisions?

## Psychological Explanations for Poor Weather-Related Decision Making

It is possible that at least some of the public's non-optimal decision making could be attributed to errors in cognition. Much of the research on cognitive limitations mentioned earlier is applicable in weather-related decision making. For instance, poor decision making might result from a typical weather-related problem structure: the cost-loss structure. In situations of weather-related risk, people are often faced with the choice between taking precautionary action, like evacuating, which incurs a cost, and not taking precautionary action, which involves no cost but incurs a potential loss. These cost-loss decision scenarios (Thompson, 1952) are framed entirely in terms of losses. Prospect theory (Kahneman & Tversky, 1979) asserts that people tend to be risk averse when faced with possible gains, a finding well known in much past research (Wu, Zhang, & Gonzalez, 2004; Kahneman, 2011), but they tend to be risk seeking when faced with possible losses (although the pattern is reversed for low probabilities; Tversky & Kahneman, 1992). This is a framing effect: responses differ depending on how the question is framed (Tversky & Kahneman, 1981). In the classic Asian disease problem (Kahneman & Tversky, 1984), participants were told to imagine that the U.S. was preparing for an outbreak of an Asian disease that was expected to kill 600 people. Participants had to choose between implementing Program A, which would result in 200 people being saved, or Program B, which would result in a one-third probability that all 600 people would be saved and a two-thirds probability that no people would be saved. Participants were overwhelmingly risk averse, preferring the sure saving of 200 lives over the possible saving of no lives (72% to 28%, respectively). Another group of participants was presented with the same problem, but they had to choose between implementing Program C, which would result in 400 people dying, or Program D, which would result in a one-third probability that nobody would die and a two-thirds probability that 600 people would die.

Participants were overwhelmingly risk seeking, preferring the chance of preventing all deaths to the sure death of 400 (78% to 22%, respectively). The decision options between the two versions of the problem were mathematically identical, i.e., they had the same expected value, the theoretical value of the decision obtained by multiplying the possible outcome by its likelihood of occurrence (von Neumann & Morgenstern, 1944). But because the former was framed in terms of gains (lives saved relative to a reference point of all lives being lost) and the latter in terms of losses (lives lost relative to a reference point of all lives being saved), participants perceived the problem differently (Kahneman & Tversky, 1984): The value of saving 200 lives with certainty was greater than the value of possibly saving 600 lives, and oppositely, the value of letting 400 people die with certainty was less than the value of possibly letting no people die. Critically, this sort of framing effect could account for people's decisions to not heed weather warnings: Faced with the cost of precautionary action and the potential loss associated with not taking action, people might be inclined toward seeking risk, i.e., taking the gamble and not heeding the warning.

Additionally, the perceived likelihood of an event often does not match its objective probability, and misperception of the likelihood of low-probability events can lead to risky decisions. Importantly, in many situations, people tend to underestimate low probabilities. Probabilities near zero are sometimes interpreted as representing no risk and might therefore be ignored (Kahneman, 2011; Lipkus, 2007). Prospect theory (Kahneman & Tversky, 1979) asserts that people tend to overweight low probabilities, but such overweighting usually occurs in situations in which a person chooses whether to accept a gamble or not, a *single* trial in which the probability of the possible outcome is *described*. This experience might not well match the experience of repeated instances of weather threats. Evidence suggests that when the occurrence

of probabilistic events is experienced rather than merely described, the perceived likelihood decreases ("description-experience gap," Hertwig et al., 2004; Erev & Barron, 2005), perhaps especially in cases of repeated false alarms, situations in which a weather event that has been warned about does not occur (Breznitz, 1985). Therefore, residents in an approaching storm's path might underestimate the likelihood of the storm because experience has led them to regard such a rare event as unlikely to occur. Additionally, there is evidence that people overestimate the costs of precautionary action (Blendon, 2008; Cutter & Smith, 2009), as well, making the inclination toward risk seeking even more pronounced. Cognitive factors like availability (Tversky & Kahneman, 1974) might make people consider all the precautions they took to prepare for a storm, for example, thus leading them to think that the stated probability of the storm is not applicable to them (Fischhoff et al., 1993). Evidence suggests people might "decode" forecasts on the belief that forecasters have exaggerated the risks or have factored in the storms' severity with the probability assessments (Fischhoff, 1995; Patt & Schrag, 2003; Windschitl & Weber, 1999). Thus, in many circumstances, it seems possible that people fail to take precautionary action for low-probability events because they underestimate the likelihood that the events will happen.

In some situations, people might be biased toward risk aversion. Indeed, not all weather-related decisions are concerned strictly with losses. Some decisions made under uncertainty, like farmers' choice of crops when future rainfall is uncertain, entail cost *and* revenue. This choice structure is captured in mixed gambles, decision alternatives in which both a gain and a loss are possible (Kahneman, 2011). Ample research has explored gains-losses framing effects; less is understood about mixed gambles. Mixed gambles have often been explored by presenting different gambles to participants and asking if they would accept the gambles or not (Kahneman,

2011). This research suggests that the disutility of a loss is generally about twice the utility of a same-sized gain (e.g., Tversky & Kahneman, 1992; Hastie & Dawes, 2010). Therefore, a gain of $200 is approximately as pleasant as a loss of $100 is unpleasant, and many people would choose to accept such a gamble with even odds of winning (i.e., 50% chance to win $200, 50% chance to lose $100) (Kahneman, 2011). A situation in which people must choose to accept or reject a given mixed gamble is different from one in which people must choose between a mixed gamble and a sure alternative (see Ert & Erev, 2008), a situation that might be more common in a real-world weather- or climate-related decision. Most evidence, however, suggests that people faced with the latter choice will demonstrate risk-averse decision making (e.g., Birnbaum & Bahra, 2007) because of loss aversion: The mere possibility of losing money, present in the mixed gamble, will make people risk averse (e.g., Payne, 2005). In other words, in some real-world weather-related decision scenarios, people might be biased toward making non-optimal decisions by taking precautionary action when doing so is not necessary.

Many factors that affect people's decision making when faced with weather-related uncertainty can be understood in terms of the "two-systems view." A critically important contribution of cognitive psychology research is the theoretical concept of two systems of cognitive processing (Stanovich & West, 2002; Sloman, 1996; Kahneman, 2011). The two-systems view posits that there are two modes of thought. System 1 is fast, automatic, effortless, associative, parallel, and emotional, and System 2 is slow, deliberate, effortful, rule-based, serial, and neutral (Kahneman, 2003). The speed and automaticity of System 1 guides most of our behavior, sending impressions to System 2, which endorses those impressions as judgments (Kahneman, 2011). Like with heuristic judgment, reliance on System 1's impressions is usually effective, but it is prone to systematic errors (Kahneman, 2011). System 2 is capable of analytical

reasoning, but System 2 is often not engaged due to cognitive constraints, leaving the more intuitive, heuristic judgments of System 1 to guide decisions. This has direct implications for how people make certain weather-related decisions. For instance, a common feature of many extreme weather situations is limited time to make a decision. Some evidence suggests that inducing cognitive constraints, such as making decisions under time pressure, distorts our perception of risk: Whereas risks and benefits are said to be positively related (i.e., generally, as the risk associated with something increases, so does its benefit), people tend to evaluate risks and benefits as inversely related, but even more so when under time pressure (Finucane et al., 2000). The theoretical explanation of this is that in the two-systems view, in decision making people rely first on automatic System 1 processes, moving to more analytical System 2 processes if time permits or if people sense an error in their intuitive judgment. Introducing cognitive load (e.g., time pressure) to a decision task thereby limits people's analytical processing and increases their reliance on intuition, which very well might influence their perception of risk and potentially lead to errors in judgment. Even if System 2 forces are engaged, they do not guarantee optimal, normative decision making, either, because System 2 has limited capacity (Baddeley, 1992).

The two-systems view might explain people's weather-related decision making. For example, determining what to do when faced with a weather-related risk might be stressful and difficult: When faced with a difficult question, e.g., Should I evacuate my house?, people might answer an easier question, e.g., Do I want to leave my house?, instead. This attribute substitution (Kahneman, 2003) is a System 1 process: The ease or convenience of evacuation is a heuristic attribute that is substituted for the target attribute of the sensibleness of evacuation. Generally, people might be representing uncertainty in two different ways, one based on automatic,

affective, intuitive reaction to it (System 1), and the other based on deliberate, calculated analysis of it (System 2). Indeed, Slovic et al. (2004) note that "people base their judgments of an activity… not only on what they *think* about it but also on how they *feel* about it" (pg. 5), suggesting parallel thought processes. Perhaps in some situations, intuitive judgment seems adequate and is not updated by more effortful analysis (Kahneman & Klein, 2009).

Additionally, the decisions people make are almost certainly affected by their trust in the forecasts. A person's decision to heed or not heed an evacuation order for an approaching hurricane – potentially a life or death situation – might ultimately be determined by the degree to which they trust the source of the evacuation order. The preservation of trust is of critical importance in the context of getting people to comply with weather warnings, because trust is significantly more easily lost than gained (Kramer, 1999; Slovic, 1999). Experimental evidence suggests that events that destroy trust tend to carry greater weight in trust judgments than events that create trust do, and trust-destroying news tends to be perceived as more credible than trust-building news (Slovic, 1993; Slovic, 1999). Emergency managers are highly aware of the importance of maintaining the trust of the people they serve, and as such, of particular practical interest in the present research is exploring ways to increase or preserve users' trust in weather forecasts. Furthermore, past research has not directly measured participants' trust in forecasts with different expressions of uncertainty.

Numerous factors could affect the trust people have in forecasts. The present research will explore two of these. First, forecast error, the difference between a forecasted event and the outcome, might affect trust. Some research suggests that people tend toward inaction in the face of variability: As outcome variability increases, people tend to take action less (Erev & Barron, 2005). Ample research on trust in automated decision support systems reveals that trust declines

when the systems make errors or are unreliable (Kramer, 1999; Seong & Bisantz, 2008), but I am unaware of any evaluation of trust over a systematic manipulation of weather forecast error in an experimental setting. Clearly, the popular narrative is that forecasters are never accurate (Silver, 2012b), but it is not clear how forecast error alone affects trust. Second, experience with false alarms, situations in which a forecasted event fails to occur, might erode people's trust in the forecast provider (Breznitz, 1985; Roulston & Smith, 2004). Some research suggests that the false alarm effect operates by way of eroding trust in the system providing the information (Bliss & Fallon, 2006; Bostrom & Lofstedt, 2003) and that repeatedly experiencing false alarms adversely affects people's response to warnings (Dow & Cutter, 1998). Others, however, claim that emergency managers' concern over public response to false alarms is unjustified (Sorensen, 2000) and that trust in general is fairly durable and is of paramount importance to confidence in forecasts (Earle & Siegrist, 2006). Thus, evidence of false alarm effects is mixed and there has not yet been a controlled experimental test of the false alarm effect in weather-related decision making.

In summary, people are sometimes biased in their understanding of weather-related uncertainty, and past research on human cognition might be able to explain why certain patterns of behavior are consistently observed, like failing to take adequate precautionary action when threatened by a storm. It is possible, though, that effective communication of weather uncertainty that takes into account the sorts of decision biases and errors above might lead people to make better decisions.

## Communicating Weather Uncertainty

Effective communication to the public is critically important in promoting good decision making about weather-related risks, and it must take into account cognitive limitations. With all the advances in calculating weather uncertainty, and given that weather is inherently uncertain, effective communication of uncertainty information would theoretically benefit forecast users. However, there is ample debate about whether or not uncertainty estimates should be communicated to the public, with the exception of the probability of precipitation, which has been included in National Weather Service forecasts since 1965 (Murphy et al., 1980). Some evidence suggests that providing uncertainty estimates enhances the credibility and trustworthiness of the source, whereas other evidence suggests the opposite (see review in Miles & Frewer et al., 2003). In many types of weather situations, such as in the event of extreme weather, probabilistic uncertainty estimates are often not communicated to the public for fear that the public will not understand them or interpret them as evidence of incompetence on the part of the forecasters (Frewer et al., 2003). Still, it is unclear if forecast users actually do not understand the uncertainty information or if scientists simply doubt they do.

There are numerous different ways that weather uncertainty can be communicated to the public, each with strengths and weaknesses. Six basic uncertainty formats are discussed below.

One common way to express uncertainty is with *numeric probability*, such as in the expression, "There is a 20% chance of flooding tomorrow." Numeric probabilities, such as percentages and classical probabilities (0 to 1), are precise and unambiguous (Murphy et al., 1980), they are easily convertible to different metrics, they carry a sense of scientific credibility, and they all use the same denominator (Lipkus, 2007), making them especially easy to compare

to each other (Cuite et al., 2008), which is often necessary when evaluating multiple risks. Despite the fact that including probabilistic information in weather forecasts is at least a 200-year-old idea (Murphy, 1998), such information has largely been left out of weather forecasts, even though the claim that users do not understand the information might be empirically unfounded (Murphy, 1980). Past research suggests that probability misinterpretation is much less common than *event* misinterpretation, e.g., misinterpreting a point forecast as an area forecast (Murphy, 1980), meaning that probabilistic uncertainty information itself is not necessarily the problem (Murphy, 1991). Inclusion of probability estimates could only serve to reduce misinterpretation of the weather events themselves (Murphy, 1980). Furthermore, forecasting errors, such as "overforecasting" (forecasters' bias toward false alarms over misses, as the latter are perceived as more serious than the former), are inherent in categorical forecasts and could be avoided by including precise numeric uncertainty estimates (Murphy, 1991). The needs of different specific user groups vary, such that inclusion of weather uncertainty estimates would theoretically cater to diverse users' risk tolerances (Joslyn & Savelli, 2010; Murphy & Winkler, 1979). However, much attention has focused on the qualities of percentages that are potentially confusing to people, such as the fact that they do not specify the reference class (Gigerenzer & Edwards, 2003). For example, the phrase "30% chance of rain" does not specify the class of events to which the 30% chance refers: The phrase could be interpreted to mean 30% of the time, 30% of the area, or on 30% of days like tomorrow (Gigerenzer & Edwards, 2003). Similarly, not specifying the reference class makes probabilities difficult to interpret for single events (Brase, 2002; Gigerenzer & Hoffrage, 1995). Additionally, for low-probability events for which precautionary action is necessary, explicit probability estimates might be underweighted because

they might not generate sufficient concern for decision makers to take precautionary action (Baker, 1995; Roulston & Smith, 2004).

An alternative to numeric probability is an expression of *frequency*, for example, "One in five times the conditions are like this, there will be flooding tomorrow." Natural frequencies, which correspond to the way in which outcomes are naturally observed (Gigerenzer & Hoffrage, 1995), are theoretically easier to understand than probabilities (Hoffrage et al., 2002; Hoffrage & Gigerenzer, 1998; Lipkus, 2007) because the reference class is made explicit (Gigerenzer & Edwards, 2003) and as such are easier to understand than single-event probabilities (Brase, 2002). Additionally, frequency expressions might help prompt action when it is required at low probabilities. Research suggests that for low-likelihood events, a frequency format is perceived as expressing greater likelihood than the equivalent probability format, perhaps because frequencies allow people to vividly imagine the cases or events being referred to (Slovic et al., 2004). In one study, some clinicians were told that a psychiatric patient, Mr. Jones, had a 10% chance of committing an act of violence toward others, whereas other clinicians were told that 10 out of 100 patients like Mr. Jones would commit an act of violence toward others. Clinicians shown the frequency format (10 out of 100 patients) rated Mr. Jones as significantly more likely to commit a violent act than did clinicians shown the probability format (Slovic et al., 2000). Therefore, this sort of frequency format enhancement could be helpful to risk communicators if they are trying to promote taking precautionary action for low-probability events, but the possibility that risk will be overweighted should be considered. Additionally, frequency expressions using different denominators can be difficult to compare, especially if a 1-in-n format is used (Cuite et al., 2008), and frequency expressions are subject to other types of biased interpretation, like the ratio bias: Expressing frequency using a ratio of two smaller numbers

(e.g., a one in five chance) results in a lower perception of likelihood than when an equivalent ratio of two larger numbers is used (a 10 in 50 chance) (Lipkus, 2007). Similarly, in one experiment, participants rated cancer that "kills 1,286 out of 10,000 people" as riskier than cancer that "kills 24.14 out of 100 people" (Yamagishi, 1997).

More complex are expressions of *relative risk*, such as, "Compared to a typical day, the chances are 1.5 times greater that there will be flooding tomorrow." Expressions of relative risk, such as odds ratios, have been used to highlight increases in risk rather than expressing the absolute risk, which might be very low (Lipkus, 2007). In one study, when participants were faced with a low-probability event, such as a tire blow-out, they were willing to pay more for products, e.g., safer tires, out of precaution when the risk was expressed in relative rather than absolute terms (Stone et al., 1994). In another experiment, medical patients' perception of benefits associated with different medications was tested, with a direct comparison of absolute and relative benefits (Malenka et al., 1993). Significantly more participants chose medications whose benefits were described in relative than in absolute terms, because the relative expression "magnified" the medication's benefits (Malenka et al., 1993). In weather forecasts, an odds ratio forecast format would therefore theoretically promote precautionary action for low-likelihood events, but it might also cause people to overestimate low-risk events that do not warrant taking precautionary action (Edwards et al., 2001).

Numeric estimates of uncertainty can be avoided by use of *verbal expressions*, such as, "Flooding tomorrow is *unlikely*." Concern about people's inability to understand numeric estimates like the ones above has led to use of verbal expressions of likelihood, such as "likely" and "slight chance of." Use of such expressions has been promoted as a way to avoid public misperceptions (e.g., Ancker, Senathirajah, Kukafka, & Starren, 2006). Unfortunately, verbal

expressions are vague and are subject to an enormous amount of interpretation variance between users (Murphy, 1991; Budescu, Por, & Broomell, 2012). The word "probable," for example, could be interpreted as meaning anything from about 50% to 95%, and "possible" might mean nearly *any* degree of likelihood to different people (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). Use of verbal expressions of probability, therefore, might lead to misunderstandings of likelihood that are even more pronounced than when using the probabilities themselves (Fischer & Jungermann, 1996; Mosteller & Youtz, 1990).

Further simplification of weather uncertainty is achieved through use of *categorical expressions*, such as, "There is a Flood Watch for tomorrow," or "Evacuation is not necessary." For example, the National Weather Service uses categorical expressions of risk for a number of weather events of Outlook, Advisory, Watch, and Warning. While such broad categories provide a quick summary of risk, there is concern that the terms are often misunderstood by the public and emergency managers alike (e.g., Shepherd, 2014). The category names themselves carry no clear meanings. However, some field research (e.g., Baker, 1995) suggests that explicit warnings, like evacuation notices from public officials, are the most important factor in people's decisions to take precautionary action.

Finally, numerous *combinations of the above formats* might theoretically provide users with more understandable information. Some problems of interpretation could be avoided with expressions like, "There is a 20% chance of flooding tomorrow. Evacuation is not necessary." For example, research has revealed that people understand climate change risk better when likelihood is expressed on a verbal-numeric scale rather than on a strictly verbal scale (Budescu, Por, & Broomell, 2012). Combining numeric probability estimates with categorical expressions, which are often used in weather emergencies (e.g., mandatory evacuation for a hurricane) would

theoretically be useful to people of varying risk tolerances but who also value categorical directives issued by emergency management or local government. However, there is no empirical evidence of the effectiveness of that combination.

What forecast formats are best at communicating weather uncertainty? Do different situations require different types of communication? Clearly, communicating weather-related risk to people is a challenge, and many of the methods traditionally used do not seem effective in all situations. Uncertainty can be expressed in numerous different ways, with theoretical benefits and limitations to each. Testing is essential to determine which expressions are most helpful to the public. The present research will directly test five of the above uncertainty expressions (all except verbal), with particular focus on numeric probability estimates.

**Testing Weather-Related Decision Making**

A critical first step in exploring how well people understand and make decisions with different expressions of weather uncertainty is to design controlled experiments. Systematically manipulating the type of forecast expressions shown to people and having them complete a decision task gives clear, quantifiable evidence of the effectiveness of the expressions.

A small but growing body of researchers has begun doing this sort of experimentation (e.g., Roulston et al., 2006; Patt, 2001; Nadav-Greenberg & Joslyn, 2009). Whereas the conventional question in cognitive psychological experiments has been, "How well do people make decisions with probabilistic information, compared to normative models of decision making?", the more optimistic line of questioning that has emerged instead asks, "How well do people make decisions with probabilistic information, compared to without it?" (Nadav-Greenberg & Joslyn, 2009). This latter line of research reveals that indeed, in certain contexts,

people make better decisions with probabilistic information than without it. In one experiment (Roulston et al., 2006), for example, participants played the role of a road maintenance company manager in charge of treating a town's roads with salt in winter if conditions warranted. Over a series of trials, participants were provided nighttime low temperature forecasts upon which to base their decisions. Participants whose forecasts included point forecasts augmented by estimates of freeze probability (or even just the likelihood of the temperature falling within the range given by the point forecast plus or minus the standard forecast error) made better decisions than participants with only point forecasts (Roulston et al., 2006). Other research using a similar paradigm produced similar results (Nadav-Greenberg & Joslyn, 2009). Importantly, the participants were non-experts (i.e., they did not have advanced training in atmospheric science or statistics). In another experiment (Patt, 2001), Zimbabwean farmers were able to learn the probability structure underlying a decision task. The farmers chose between two crops, maize and millet, and then spun a wheel with proportions representing varying probabilities of a wet year and a dry year, with the outcomes (represented by small cash rewards) of the farmers' crop choices being determined by whether the years were wet or dry. Results suggested that the rural farmers, with little to no formal math education, were able to make effective use of the probabilistic information, making increasingly economically optimal decisions over several rounds of play.

In some ways, it is not surprising that uncertainty estimates help, for there are theoretical benefits of their inclusion. First, research has revealed that people have reasonably good intuition about weather uncertainty (Joslyn & Savelli, 2010; Morss et al., 2008). Weather forecasts are inherently uncertain, and people know it: A lifetime of observing forecast-outcome mismatches has provided a strong intuition. Single-value forecasts imply determinism, wherein initial

conditions ultimately lead to an outcome (i.e., the outcome can be determined from the initial conditions), in no way involving probabilistic processes (Hoefer, 2008). People correctly intuit that forecasts are not deterministic. Therefore, forecasts that include uncertainty estimates quite likely seem more realistic to people, and as such, people might trust these forecasts more. Second, forecast-observation mismatches might seem less wrong to users when uncertainty estimates are included in forecasts, because the forecaster has acknowledged up front that the observed temperature (or weather event) could not be predicted with exact precision, and because whether the event occurs or not, a probabilistic prediction cannot be wrong, per se. This might also increase users' trust. Third, research has revealed variation in risk tolerances, both between people and between activities, for everyday decisions and specifically for different types of weather events (Morss et al., 2010; Joslyn & Savelli, 2010). Including uncertainty estimates in forecasts, therefore, would cater to individuals' different risk tolerances and theoretically make the forecasts more useful.

These findings are of critical practical importance: Any measure that can be taken to *improve* the general public's decision making, especially in situations of potentially dangerous extreme weather, ought to be considered, even if such a measure does not raise the public's decision making to normative levels. Moreover, it is possible that many of the numerous shortcomings in weather-related judgment and decision making discussed earlier could be overcome by including numeric uncertainty estimates. For example, whereas weather warnings communicated with verbal expressions of uncertainty might be vague and therefore misunderstood, warnings with numeric uncertainty estimates would be unambiguous and precise, and such expressions would cater to different user groups with diverse risk tolerances. Similarly, the ignoring of weather warnings because of past experience with false alarms might be reduced

by the inclusion of uncertainty estimates, because end-users could base their decisions on their own risk tolerances rather than on externally defined action thresholds (Roulston & Smith, 2004).

However, there is still very little experimental research on communicating weather-related uncertainty and how people make weather-related decisions. Many questions remain unanswered. For example, no research of which I am aware has systematically explored the effect of increased forecast error or false alarms on decision making and trust. Some past experiments that tested weather-related decision making (Roulston et al., 2006; Nadav-Greenberg & Joslyn, 2009) used experimental tasks that featured a cost-loss structure, but there have not been any experimental tests of weather-related decision making with situations involving choosing between a gain and a mixed gamble; it is unclear how people make decisions in such situations or if the beneficial effects of inclusion of uncertainty estimates generalize to such decision structures. What also remains unclear is the degree to which the benefits of inclusion of weather uncertainty estimates generalize across different real-life contexts. For example, including probability estimates might improve decisions over a wide range of probabilities, but for rare events with low probabilities, explicitly stating the probability estimate might discourage taking precautionary action. The low probability of an event, even an extreme one for which precautionary action is necessary at low probabilities, might be discounted or ignored altogether, resulting in a potentially costly decision error. In such situations, providing an estimate of relative risk might be preferable, or perhaps giving categorical decision advice would facilitate decision making.

Finally, the effect of decision feedback requires further exploration, as feedback potentially plays a critical role in understanding of weather-related uncertainty. There is an

expansive literature within psychology on the effects of feedback on learning (see Kluger & DeNisi, 1996). Much past research has identified decision feedback as a fundamentally important component of learning (e.g., Einhorn & Hogarth, 1981). Research on expertise and expert intuition suggests that the opportunity to learn an environment, including clear and immediate feedback from decisions and actions, is essential for the development of skilled intuitions (Kahneman & Klein, 2009) and that the availability of feedback is a predictor of good performance among experts (Shanteau, 1992). Experimental evidence suggests that both individuals and groups made better decisions in a hypothetical staff hiring process when presented with outcome feedback than when feedback was withheld (Tindale, 1989). In another study (Diehl & Sterman, 1995), participants played a generic stock adjustment task in which they managed the inventory of a product by increasing production when inventory was low. Feedback complexity was systematically manipulated, and as feedback became more difficult to understand, participants' task performance decreased. These experiments demonstrate the importance and benefit of feedback in task performance. Other research, however, suggests that past findings are decidedly more mixed and often contradictory (Jacoby et al., 1984; Kluger & DeNisi, 1996), although the inconsistency of past results is often related to the type of feedback provided, e.g., positive or negative, rather than the presence or absence of feedback. In one experiment (Weber, 2003), groups of participants played a competitive game with multiple rounds in which the outcome of each round theoretically determined the decision each participant would make in the subsequent round. For some groups, outcome feedback was given after each round, and for others, no feedback was given until the end of the experiment. Results showed that participants' strategies approached the game's optimal strategy whether they received feedback or not, suggesting that learning occurred merely by gaining experience with

the environment (Weber, 2003). For decisions that concern far-future events, like climate-related decisions, outcome feedback might not be available for significant amounts of time. Past research generally suggests that the lack of immediate feedback would be detrimental to the quality of climate-related decisions, but I am unaware of any research that has systematically explored this in a controlled setting. Furthermore, it is unclear if providing numeric probability estimates would lead to better decisions and higher trust among decision makers, and if such effects are contingent upon immediate decision feedback being provided. Uncertainty estimates might enhance decision quality and trust in weather forecasts because those estimates account for the potential mismatch between forecasted weather events and the actual weather conditions observed soon thereafter. Uncertainty estimates might not have the same effect for climate projections, however, as the actual climate conditions will not be observed immediately.

**The Present Research**

The small body of research that reveals benefits of including uncertainty estimates in forecasts is an encouraging development in cognitive psychology research and stands in contrast to the overarching negative assessment of people's ability to reason with probability. However, critical questions remain: What is the extent of the benefit of inclusion of probabilistic estimates? Are there situations in which providing probabilities is not beneficial to decision makers? Are there boundary circumstances? More specifically:

1. High forecast error: What is the effect of forecast error on decision quality? Do uncertainty estimates still improve decision quality if forecast error is increased? (Chapter 2)

2. Alternatives to uncertainty estimates: Are there ways to communicate weather-related uncertainty that are superior to inclusion of uncertainty estimates, like providing explicit decision advice or adjusting false alarm rates? What is the effect of false alarms on weather-related decision making? (Chapters 2 and 3)
3. Rare, extreme weather events: How should uncertainty be communicated in situations of rare weather events? (Chapter 4)
4. Distant-future events: Does inclusion of uncertainty estimates help decision makers with far-future decisions, including decisions without feedback? Do uncertainty estimates influence perception of climate change risk? (Chapters 5 and 6)

The present dissertation research seeks to answer these questions.

The research covered in this dissertation is applied research: I have started with real-life situations and problems, identified the components of those problems that are theoretically believed to drive the outcomes (in this case, human behavior), and carefully designed laboratory experiments that systematically vary the components of interest and hold everything else constant, to the greatest extent possible. While some ecological validity is compromised in this process, the experimental control I exert allows for more readily interpretable findings, including conclusions about causal relationships, which are often extremely difficult or impossible to make in field research. Importantly, much of this applied research is exploratory. I have not begun with basic theoretical concepts and linked them to the application of interest. Rather, I have begun with the application and attempted to use theoretical concepts to inform the manipulations and give structure to the results. In some cases, fitting the findings with theoretical explanations has not been straightforward. This is not a weakness of the research but simply an inherent aspect of

it, and it strongly suggests that further research should continue exploration of this and other applications in order to provide a sturdy theoretical foundation.

In the present context, "decision making" is defined in the manner outlined by Hastie & Dawes (2010):

> *"A decision in scientific decision theory terms is a response to a situation that is composed of three parts: (a) There is more than one possible course of action under consideration…; (b) the decision maker can form expectations concerning future events and outcomes following from each course of action, expectations that are often described in terms of probabilities or degrees of confidence…; and (c) consequences, associated with the possible outcomes, that can be assessed on an evaluative continuum reflecting personal values and current goals." (p. 25-26)*

Additionally, Hastie & Dawes (2010) characterize decisions as "deliberate, conscious accomplishments" (p. 26) and limit the study of them to their immediate consequences, qualities that characterize how decisions will be explored in the present research.

Finally, "risk" and "uncertainty" are often used interchangeably in everyday speech. The terms are sometimes used interchangeably in the present research, as well, but in the present context, I will generally be referring to Frank Knight's (see Hau, Pleskac, & Hertwig, 2010) characterization of "uncertainty": Whereas "risk" concerns situations with well specified, *a priori* outcome probabilities, like rolls of dice or flips of a coin, "uncertainty" concerns outcomes determined by natural events for which a probability can only be estimated (Trepel, Fox, & Poldrack, 2005; Hau, Pleskac, & Hertwig, 2010). Additionally, the uncertainty explored in the present research – specifically, how weather- and climate-related uncertainty is communicated and perceived – is "first-order uncertainty," which characterizes predictions about future events as uncertain due to the fact that the events are probabilistic, in contrast to "second-order uncertainty," which characterizes uncertainty about the event probabilities themselves (Bach, Hulme, Penny, & Dolan, 2011).

Several of the experiments to follow feature a basic decision-making problem that fits Hastie & Dawes's (2010) characterization and matches the structure of the analogous real-world decision scenarios upon which they were designed. In the experiments, participants make a series of decisions in which they are faced with a potential threat. In some experiments, decision outcomes are framed entirely as losses, matching the analogous real-world situation: Participants must decide whether they ought to spend a limited resource on precautionary action, thereby averting the risk, or take a gamble, saving on the precautionary cost but subjecting themselves to the possibility of a relatively severe penalty. As the costs and potential losses in these cost-loss scenarios are quantified, cost-loss ratios (Thompson, 1952) can be calculated. These values – the cost divided by the potential loss – serve as important thresholds that clearly identify economically optimal decisions. They can be used to develop output for decision advice (see Chapters 2 and 3) and serve as standards against which participants' decisions can be compared. In another experiment, a mixed gamble is used and decision outcomes are framed in terms of gains and losses, again reflecting the real-world situation upon which it is based. Most experiments also feature repeated-trials designs, allowing participants to learn the relationship between stimuli and outcomes, where applicable, and providing a richer set of data from which to draw conclusions. The primary overarching research question is how inclusion of estimates of weather uncertainty, expressed in various ways, affects decision making, so all experiments include a manipulation of forecast expression as a key component. All experiments and questionnaires include direct and indirect measures of participants' trust. Most importantly and generally, all of the research that follows concerns real-life weather- and climate-related problems.

The following experiments were designed primarily to explore real-world weather-related decision problems in a controlled setting and to answer specific questions about how best to communicate weather uncertainty estimates in different situations. However, the experiments also allow for exploration of the applicability of the basic theoretical material reviewed earlier, such as the tendency to seek risk in situations of losses. Altogether, the research will contribute both to the applied research literature and to the basic psychology literature, with implications both about how real-world risk communication can be improved and about how people make decisions when faced with uncertainty.

## Chapter 2: Decision Advice and Forecast Error

### Background

In the growing body of recent research that shows an advantage to decision makers who have weather uncertainty estimates (e.g., Nadav-Greenberg & Joslyn, 2009; Roulston et al., 2006), two important aspects of real-world situations involving communication of weather-related risk are left unexplored: 1) forecast error and 2) the need for precautionary action at low probabilities.

A critical aspect of the communication of weather-related risk to vulnerable residents threatened by extreme weather is forecast error, the difference between what is forecasted and what actually occurs. Weather warnings must be communicated long enough in advance to allow residents to take precautionary action. However, longer lead times are inherently associated with lower forecast accuracy (e.g., Kootval, 2008; Brooks, Witt, & Eilts, 1997). A careful balance, therefore, must be found between the benefit of giving early warning and the detriment caused by high-error forecasts. It is possible that high forecast error brings about user skepticism and distrust, potentially leading to future instances of ignoring forecasts and weather warnings, but the relationship between forecast error and trust has not been systematically tested. In the few past experiments on the effects of inclusion of uncertainty estimates (Roulston et al., 2006; Nadav-Greenberg & Joslyn, 2009), participants made better decisions overall when provided with uncertainty estimates. However, in those experiments, forecast error was not systematically manipulated. Therefore, it remains unclear what effect, if any, increased forecast error has on users' decisions and if any such effect could be attenuated by inclusion of uncertainty estimates. Including uncertainty estimates might make forecast-observation mismatches, even large

mismatches in the case of high forecast error, seem less inaccurate, as the uncertainty estimate highlights the probabilistic nature of the observation. But the effect is not known, as the effect of forecast error on trust – and the potential benefit of including uncertainty estimates – has not yet been explored in a controlled setting.

Additionally, because of the relatively large potential loss associated with inaction in the event of extreme weather (e.g., loss of life), precautionary action often must be taken at low probabilities, probabilities so low that they might not generate sufficient concern on the part of decision makers (Baker, 1995; Roulston & Smith, 2004). As such, when precautionary action is required even at low probabilities, communicating uncertainty estimates, e.g., probability of a specified storm surge, might perversely mislead vulnerable residents toward *not* taking precautionary action because doing so would often result in false alarms (Roulston & Smith, 2004). Probabilities that are already low might still be underestimated because repeated experience with non-occurrence of low-probability events suggests to decision makers that the events are less likely than they actually are (Erev & Barron, 2005; Weber, 2006). Indeed, as was noted in Nadav-Greenberg & Joslyn (2009), participants who were shown freeze probability estimates did not benefit from explicit probability information at low probabilities for which precautionary action was the economically optimal course of action. Instead, people might benefit from being given explicit advice about what to do (i.e., take precautionary action or not), without being told the actual probability of the event. A categorical risk expression would clearly advise what to do, and people would not be able to underweight the low underlying event probability. This is essentially the approach taken in many real-life weather warning situations, in which emergency managers or other authority figures issue warnings, e.g., evacuation orders, without explanation of the underlying probabilities of the events being warned about.

However, this might not be an effective approach, either. Field research provides substantial evidence that a large proportion of populations threatened with severe weather often do not heed weather warnings. For instance, in September 2008, as Hurricane Ike approached the Gulf Coast of the United States, fully 40% of residents did not comply with mandatory evacuation orders, despite being warned of "certain death" (McKinley & Urbina, 2008). Similarly, of residents under mandatory evacuation for Andrew and Hugo, both Category 4 hurricanes, only 42% heeded evacuation orders (Riad et al., 1999), and only 64% evacuated for Hurricane Floyd (Dow & Cutter, 2000). An official report commissioned by the City of New York found that only 33% of interview respondents living in low-lying Zone A in New York City evacuated when Hurricane Sandy approached (Gibbs & Holloway, 2013). -During the period from 2009 to 2011, evidence suggests that people inside the warning polygon were no more likely to seek shelter from a possible tornado than were people outside the polygon (Nagele and Trainor, 2012). Evidence also exists that people do not take adequate precautionary action when flood warnings are issued (Perry, 1983; Gruntfest et al.,1978; Parker, Priest & Tapsell, 2009). Clearly, explicit advice does not always prompt people to take action.

Still, the effect of explicit decision advice – in comparison to the effect of uncertainty estimates – has not yet been explored systematically in a controlled setting. The poor compliance observed in the field might have been due to a number of different factors other than the weather warning itself. It remains unclear if including decision advice with forecasts will lead to better decisions, particularly for low-probability events, than including uncertainty estimates, or if people will not heed the decision advice, as has been demonstrated time and again in actual weather emergencies.

Another alternative to getting people to take precautionary action at low probabilities is with a frequency expression of uncertainty. For example, a 15% chance of an event could be described as 15 instances out of 100 of the event occurring, or in 15 out of 100 "situations like this." Frequency formats might be easier to understand because they are more naturally aligned with how people learn likelihoods (e.g., Hoffrage et al., 2002) and potentially easier to understand than single-event probabilities (Brase, 2002), although people might overweight a risk when it is expressed as a frequency compared to when it is expressed as a probability. It is unclear, though, if a similar probability-frequency gap is demonstrated in weather-related decision making. To explore the influence on decision making of frequency expressions of weather uncertainty compared to probability expressions, the present experiment will test both a probability and a frequency expression.

In the following three experiments, I systematically tested two factors of critical importance in real-world weather-related decision making: 1) the effect of forecast error on decision making, and if inclusion of uncertainty estimates has different effects on low- versus high-error forecasts; and 2) the effect of different forecast formats on decision making, including expressions of probability, frequency, and explicit decision advice, and if decision quality for low-probability events improves or declines with decision advice compared to forecasts that include uncertainty estimates. I also tested a hybrid: decision advice combined with uncertainty estimates. In the experiment, participants completed a computerized task in which they played the role of a manager of a road maintenance company and used low temperature forecasts to make a series of decisions about whether or not to take precautionary action to prevent icy conditions on local roads.

**Road Salting Experimental Paradigm**

Many of my hypotheses were tested using an experimental paradigm called the road salting task. The basic experimental task was originally developed by Roulston et al. (2006) to test the effect of the inclusion of uncertainty estimates in weather forecasts. The task was modified for my lab's use (see Nadav-Greenberg & Joslyn, 2009), and I have created multiple versions of it over several years in order to explore specific questions of interest.

The task is a computer-based game but is administered in person to participants in a computer lab. In the task, participants assume the role of a manager of a road maintenance company in contract with an American town to keep the town's roads ice-free during winter months. Over a series of trials, each representing one day, participants face a cost-loss scenario in which they must decide if they should pay to apply salt brine to the roads, which prevents ice from forming, or withhold the salt brine and risk a penalty to cover damages to the town's infrastructure caused by auto accidents if a freezing temperature is observed. Participants are given a limited starting budget, so decisions must be made carefully. Treating the roads incurs a cost, whereas withholding treatment does not, but if a freezing temperature is observed and the roads have not been treated, the road maintenance company is charged a substantial fine. To help participants make this decision, all participants are shown a forecast for the overnight low temperature on each trial. Based on the forecast, if participants believe the overnight low temperature will fall to 32°F or lower, they should choose to apply salt treatment; if they believe the temperature will be 33°F or higher, they should choose to withhold salt treatment. Participants are also asked to give 1) a trust rating, indicating how much they trust the forecasts to help them make their decisions, during each trial and after each block of 30 trials (representing a month); and 2) their own estimate of the overnight low temperature. Participants are instructed

that their goal is to maximize their budgets. To encourage them to do so, participants are paid real cash commensurate with their performance at the game. For example, a participant who makes economically optimal decisions would be rewarded about $5 to $10.

The road salting task is a substantially simplified version of its real-world analogue. For instance, road maintenance managers' salting decisions are based on numerous factors in addition to overnight low temperature (Roulston et al., 2006), and certainly the temperature merely being 32°F or below would not necessarily mean that the roads would become icy and cause a set amount of damage. Still, the task preserves important characteristics of many real-world decisions: a binary choice between costly precautionary action and a gamble that *might* be substantially more costly, with a probabilistic outcome that has real consequences for the decision maker.

**Experiment 1: Decision Advice and Forecast Error**

In Experiment 1, forecast error was systematically manipulated. Additionally, some participants were shown forecasts that included uncertainty estimates, and other participants were shown forecasts that included explicit decision advice. It was hypothesized that increased forecast error would result in inferior decisions and reduced trust, simply because the predictive quality of high-error forecasts would be lower than that of low-error forecasts; and that uncertainty estimates would attenuate the reduction in decision quality and trust by making the high-error forecasts seem less inaccurate. It was unclear what effect explicit advice would have on decision making.

## Method

Participants. Three hundred and four participants (50% female) took part in the experiment. Age ranged from 18 to 50 years ($M$ = 19.5). All participants were recruited from introductory psychology courses at the University of Washington and were awarded a small amount of academic credit for participating. Participants signed up voluntarily and had a chance to win a small cash prize.

Apparatus. The experiment was programmed with Microsoft Excel Visual Basic and was administered via Excel on standard desktop computers.

Procedure. Participants arrived in groups of 1 to 12 at pre-organized times in a computer lab. After participants read a consent form and agreed to participate, the experimenter presented task instructions and a practice trial while participants followed along on their computers. Following an opportunity for questions, the task began. Participants completed the task individually. Upon completion of the task, participants answered follow-up questions concerning what strategies they used to make their decisions and if they experienced any problems with the program. All participants who showed up for the experiment were awarded credit.

The experiment used the road salting paradigm. Participants were presented with a series of 120 trials, representing 120 days (broken down into four 30-day months) in winter. On each trial, using the forecast for overnight low temperature, participants first selected a trust rating on a five-point scale (from "very little" to "very much"), indicating how much they trusted the forecast to help them make an informed decision. Then they had to decide between two options: 1) spend $1,000 on salt treatment to prevent icy conditions on the town's roads, or 2) spend nothing and withhold treatment but risk a $6,000 penalty if a freezing temperature were

observed. Salt treatment cost $1,000 whether a freezing temperature was observed or not; withholding treatment cost nothing but resulted in the $6,000 penalty in the event of a freezing temperature. See Table 2.1. Finally, participants entered their own estimate for the nighttime low temperature. After this, they were shown feedback, the observed nighttime low temperature. By clicking a "next" button, they proceeded to the next trial. After each block of 30 trials, they gave another trust rating (on the same scale as before), indicating how much they generally trusted the forecasts over the past virtual month. After the four virtual months, participants completed the experiment by answering some follow-up questions soliciting their strategies, comments, and any problems they experienced with the task. Participants were dismissed as they finished. Those who earned a cash prize (see below) were paid and dismissed. All participants received credit.

Table 2.1. Task cost structure: costs and losses associated with decisions to salt or not salt.

| | **Outcome** | |
|---|---|---|
| | **Freezing temperature (≤ 32°F)** | **Non-freezing temperature (> 32°F)** |
| **Salt** | $1,000 | $1,000 |
| **Not Salt** | $6,000 | $0 |

Note: This table shows costs and losses, e.g., a cost of $1,000 is equal to -$1,000.

Participants were given a monthly budget of $36,000, such that if they salted on every trial to avoid risk of penalty, they would finish the task with a balance of $24,000. As such, to incentivize participants to try their best at the task, they were paid $1 in cash for every $1,000 over $24,000 in their virtual balances at the end of the experiment. For example, a participant finishing the task with $29,000 would be given $5.

The ratio of the $1,000 salt treatment cost to the $6,000 potential penalty is one-sixth, or 16.7%. That probability, or cost-loss ratio, represents the exact point (in this case, freeze probability) at which the long-run payoff of applying salt treatment would break even with the

long-run payoff of withholding salt treatment. Given that the road salting task used only whole-number probabilities of freezing, this means that in order to maximize their budgets, participants ought to have applied salt treatment whenever the probability of an overnight freezing temperature was 17% or greater, and they ought to have withheld salt treatment whenever the probability was less than 17%. Following that rule would, in the long run, lead to an economically optimal result.

Stimuli. Forecasts and freeze probabilities were taken from climatological records obtained from the Department of Atmospheric Sciences at the University of Washington. A sample of 60 forecasts from winter months in Yakima and Spokane, WA, was selected. Each forecast in the forecast set was treated as the mean of a normal distribution of possible temperatures. Using the associated probability of a freezing temperature (32°F or below), I calculated standard deviations for each distribution. Observed temperatures were generated for each distribution by using Excel's random number tool, which produced values from 0 to 1, representing the area beneath the curve of each distribution and providing an observed temperature value. I created a set of 60 trials in this way. The forecast set had a mean forecast error (difference between forecasted and observed temperatures) of 3.26°F. These 60 forecasts comprised the low-error forecasts. I then created a matched set of high-error forecasts by doubling the standard deviation of each individual distribution, resulting in a new set of 60 forecasts with a mean forecast error of 6.53°F. Consequently, low-error forecasts and observations were substantially more highly correlated ($r$ =0.30) than high-error forecasts and observations ($r$ = 0.11). Additionally, the probability of freezing associated with each of the high-error forecasts was different (i.e., higher, or closer to 50%) than their low-error counterparts, and of course the observed temperatures were more extreme (i.e., positive errors became more positive, negative errors became more

negative), but the sequence of 60 single-value low-error forecasts matched the sequence of 60 single-value high-error forecasts. See Table 2.2.

Table 2.2. Forecast characteristics.

| | **Low-error forecasts** | **High-error forecasts** |
|---|---|---|
| **Mean forecast error** | 3.26°F | 6.53°F |
| **Mean freeze probability** | 24.75% (range 10-51%) | 36% (range 26-51%) |
| **Mean forecast** | 34°F (range 32-37°F) | 34°F (range 32-37°F) |
| **Mean observation** | 35°F (range 27-42°F) | 35°F (range 21-51°F) |

The sequence of forecasts and observations followed a naturalistic course, with roughly simulated weather patterns, or sub-sequences of temperatures increasing and decreasing over time. From one trial to the next, there was never a difference in observed temperature greater than 16°F, consistent with my actual weather data.

Importantly, the observed temperatures accurately reflected their associated freeze probabilities, i.e., the uncertainty estimates were reliable. With a sufficiently large set of trials, one could achieve precise calibration, e.g., freezing temperatures observed on 24% of trials in which there was a freeze probability of 24%. With only 60 base trials, however, calibration was achieved by creating bins of probabilities and fitting freeze rates within them. For example, freezing temperatures were observed on about 11% of trials in which the freeze probability was between 10% and 16%. Freeze probability and temperature forecast were extremely highly correlated, both for low-error trials ($r$ = -0.91) and high-error trials ($r$ = -0.92).

Design. The experiment used a 4 x 2 mixed-model design. There were two independent variables: forecast format, manipulated between participants, and forecast error, manipulated within participants. Forecast format had four levels: freeze probability, in which participants

were presented with overnight low temperature forecasts that included an estimate of uncertainty expressed as the percent chance of freezing temperatures (e.g., 25% chance); freeze frequency, in which the forecasts included an estimate of uncertainty expressed as the number of "days like this" out of 100 in which a freezing temperature is observed (e.g., 25 out of 100 days like this); advice, in which the single-value forecast was coupled with a decision recommendation from a Decision Support Aid (described below); and control, in which the single-value forecast alone was presented. See Table 2.3. The within-participants variable, forecast error, had two levels: low error and high error. Each level consisted of 60 trials, presented in blocks. These blocks were counter-balanced between participants: Some participants were presented the low-error block first, followed by the high-error block, while other participants had the reverse order.

Table2.3. Example forecasts for each forecast format.

| **Forecast format** | **Example forecast** |
|---|---|
| Freeze probability | "The expected nighttime low temperature is 35°F; there is a 20% chance the temperature will be 32°F or less." |
| Freeze frequency | "The expected nighttime low temperature is 35°F; 20 out of 100 days like this, the temperature will be 32°F or less." |
| Advice* | "The expected nighttime low temperature is 35°F. DSA: 'Salting is recommended under these circumstances.'" |
| Control (single-value) | "The expected nighttime low temperature is 35°F." |

*Participants in the advice condition were given additional instructions about the decision support aid, as follows: *"To help you make your decisions, your company uses the Decision Support Aid (DSA) advanced weather modeling computer system, which incorporates the most recent weather forecast available, the uncertainty involved, and the costs associated with salting or not salting, and provides you with a decision recommendation for each day's forecast."*

The Decision Support Aid (DSA) was described to participants as a tool to help the road management company make good decisions. The DSA was said to incorporate forecast information and the cost and penalty associated with treating or not treating the roads in order to provide a treatment recommendation for each trial. The DSA's recommendations were based on the cost-loss threshold of 17%: If the freeze probability on given trial were 17% or greater, the

DSA advised applying treatment, and if the freeze probability were less than 17%, the DSA advised withholding treatment. In the present experiment, participants were not explicitly aware that the DSA's recommendations were based on the cost-loss ratio, nor was the cost-loss ratio itself explicitly explained.

**Results**

I hypothesized that participants who were shown forecasts that included numeric uncertainty estimates would make better, more economically optimal decisions than participants with single-value deterministic forecasts would, and that that superior performance would be matched with higher ratings of trust in the forecasts. I did not hypothesize how advice participants would perform. I also hypothesized that all participants would do worse when forecast error increased, but that uncertainty participants' performance would be affected less.

Before analyses, I eliminated the data of any participant I suspected was not taking the task seriously. To do this, I looked at participants' temperature estimates and calculated mean standardized error scores for each condition. I removed participants whose individual error scores were two or more standard deviations above the group means. Twenty participants' data were removed, leaving 284 participants' data for analysis.

In the following analyses, subsets of the data were used. Analysis and observation suggested that many participants depleted their virtual budgets by the end of the final month (Month 4). It is difficult to determine the resulting psychological and practical effects of this happening. Some participants might have become more risk seeking, and others might have disregarded the stimuli simply to finish the task as quickly as possible. As such, for analyses in which low- and high-error trials were compared, I did not analyze data from Month 4, and

because Month 4 was matched to Month 2, I eliminated Month 2 from analysis, as well, leaving Months 1 and 3 for those analyses. Additionally, the within-participants forecast error variable was counterbalanced to account for possible order effects. Analysis revealed significant order effects. As such, except where noted, only low-error-first participants' data will be discussed. Three dependent measures will be analyzed: mean expected decision value, trust, and binary decision. Additionally, participants' responses to an open-ended follow-up question about strategy will be evaluated.

Participants with uncertainty estimates made better decisions than participants with deterministic forecasts did. This was determined by an analysis of mean expected decision value. While final balance is perhaps a more immediately compelling dependent measure, it is not a pure reflection of participants' decision quality. This is because final balance is influenced by chance. For example, a participant might choose to withhold salt treatment on a trial with 40% freeze probability – a risk-seeking error, based on the task's cost-loss ratio – but not be penalized because of a non-freezing observation. Similarly, a participant might appropriately choose to withhold treatment on a trial with 12% freeze probability but be penalized because of an unlikely freezing observation. Instead, mean expected decision value is a more accurate reflection of participants' choices, as it is irrespective of chance. For each participant, an expected decision value was calculated for each trial, either simply -$1,000 for decisions to apply salt, or, for decisions not to salt, by multiplying the potential penalty, -$6,000, by the probability of freezing on that trial. Thus, these values were the theoretical expected values (von Neumann & Morgenstern, 1944) of participants' decisions in the long run, over a sufficiently large number of trials. The mean of these expected values over all trials was calculated and used for analysis. Each participant had a single mean expected decision value. As the optimal mean expected value

differed between the low- and high-error trials (-$924 and -$1,000, respectively), I subtracted those values from participants' mean expected values, leaving difference scores (actual – optimal) for analysis.

Participants whose forecasts included uncertainty estimates had higher mean expected decision values than participants with deterministic forecasts. A mixed-model ANOVA was run on mean expected decision value, with forecast format (freeze probability, freeze frequency, advice, and control) as the between-participants independent variable and forecast error (low and high) as the within-participants independent variable. The analysis revealed a significant main effect of forecast format, $F(3, 146) = 5.83$, $p < .01$. Tukey's post hoc analysis revealed significantly smaller differences (better performance) among freeze frequency participants than either control participants, $p < .01$, or advice participants, $p = .05$, and freeze probability participants performed significantly better than control participants, $p = .01$. No other format differences reached significance. See Table 2.4.

Table 2.4. Mean expected decision value, mean monthly trust rating, proportion of risk-averse errors, and proportion of risk-seeking errors (standard deviation) by condition.

| **Forecast format** | **Mean expected decision value (difference)*** | **Mean month trust rating** | **Risk-averse error proportion** | **Risk-seeking error proportion** |
|---|---|---|---|---|
| Freeze probability | -$222.84 ($142.01) | 2.58 (.71) | .10 (.10) | .65 (.22) |
| Freeze frequency | -$204.70 ($142.17) | 2.82 (.74) | .21 (.23) | .54 (.27) |
| Advice | -$292.78 ($163.58) | 2.21 (.69) | .22 (.16) | .51 (.23) |
| Control | -$330.47 ($141.35) | 2.14 (.65) | .30 (.23) | .51 (.24) |
| *mean* | *-$261.66 ($154.84)* | *2.44 (.74)* | *.21 (.20)* | *.55 (.25)* |

*Lower absolute values signify better performance. For example, a value of -$200 indicates better performance than a value of -$300.

Additionally, participants did worse on high-error trials than on low-error trials. The mixed-model ANOVA revealed a significant main effect of forecast error, $F(1, 146) = 334.44$, $p < .01$, with significantly worse expected value differences in the high-error trials. Moreover, although participants overall did worse on high-error trials than on low-error trials, participants with uncertainty estimates (probability and frequency) demonstrated less of a deterioration in performance than participants with deterministic forecasts (advice and control) did. In other words, there was a smaller decline in decision quality due to high forecast error for uncertainty participants than for deterministic participants. This was revealed by the significant interaction, $F(3, 146) = 4.10$, $p < .01$. See Table 2.5 and Figure 2.1.

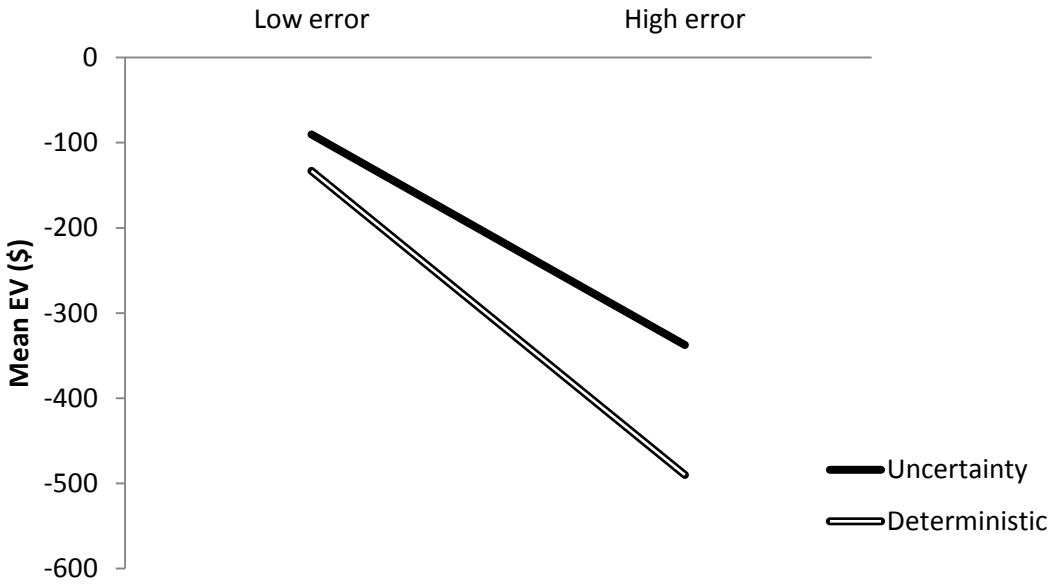
Interestingly, the effect of forecast format disappeared when the high-error trials were presented first. Uncertainty and deterministic participants performed about the same on low-error trials when those trials followed the high-error trials. Moreover, performance on low-error trials was worse when those trials were presented after high-error trials than when they were presented first, suggesting that the negative impact on decision quality of the high-error forecasts endured through the low-error forecasts. Among participants who were presented with high-error forecasts first, a mixed-model ANOVA on mean expected decision value, with forecast error as the within-participants variable and forecast format as the between-participants variable, revealed a significant main effect for forecast error, $F(1, 130) = 319.62$, $p < .01$, with significantly worse performance in the high-error trials than in the low-error trials. There was not a significant effect of forecast format, nor was there a significant interaction. (Note that this analysis used high-error-first participants, unlike the previous analyses.) See Table 2.5. An analysis directly comparing mean expected decision value on low-error trials between low-error-

first and high-error-first participants suggested that decision value was significantly lower when low-error forecasts were presented after the high-error forecasts, $t(282) = 5.43$, $p < .01$.

Table 2.5. Mean expected decision value (standard deviation) on low- and high-error trials for low-error-first and high-error-first participants by forecast format.

| | **Low-error-first participants** | | **High-error-first participants** | |
|---|---|---|---|---|
| **Forecast format** | **Low-error trials** | **High-error trials** | **Low-error trials** | **High-error trials** |
| Freeze probability | -$90.79 ($54.36) | -$354.89 ($240.48) | -$165.10 ($94.18) | -$452.48 ($190.10) |
| Freeze frequency | -$89.44 ($55.91) | -$319.96 ($245.53) | -$165.26 ($122.02) | -$400.29 ($205.02) |
| Advice | -$126.87 ($75.09) | -$458.68 ($262.99) | -$169.47 ($113.72) | -$381.82 ($201.56) |
| Control | -$139.52 ($80.03) | -$521.41 ($224.22) | -$179.05 ($104.88) | -$452.71 ($176.04) |
| *mean* | *-$111.21 ($69.91)* | *-$412.12 ($254.49)* | *-$169.48 ($108.71)* | *-$420.19 ($194.56)* |

Figure 2.1. Interaction of forecast error (low and high) and forecast format (uncertainty and deterministic). Note that freeze probability and freeze frequency were combined as "uncertainty," and advice and control were combined as "deterministic."



The mean of the end-of-month trust ratings for each participant was used for analysis, as the end-of-month ratings turned out to be comparable to the trial-by-trial ratings. Overall, participants with uncertainty estimates indicated higher levels of trust than participants with deterministic forecasts. Like with the expected decision value analysis, a mixed-model ANOVA was run on trust ratings (5-point scale, from 1-very little to 5-very much), with forecast format (freeze probability, freeze frequency, advice, and control) as the between-participants independent variable and forecast error (low and high) as the within-participants independent variable. The analysis revealed a significant main effect of forecast format, $F(3, 146) = 7.84$, $p < .01$. Tukey's post hoc analysis revealed that freeze frequency participants gave significantly higher trust ratings that either control participants ($p < .01$) or decision advice participants ($p < .01$), and freeze probability participants gave significantly higher ratings than control participants, $p = .04$. No other format differences reached significance. Again, there was a main effect of forecast error, $F(1, 146) = 99.52$, $p < .01$. Participants gave significantly higher trust ratings in the low-error trials than in the high-error trials. There was not a significant interaction. See Table 2.4.

Among participants who were shown high-error forecasts first, trust was also lower for high-error forecasts than it was for low error forecasts. Interestingly, consistent with the expected value analysis, presenting high-error forecasts first had a negative effect on trust, as low-error forecasts were rated less trustworthy when they were presented after the high-error forecasts. These effects were determined by performing a mixed-model ANOVA on trust ratings, with forecast error as the within-participants variable and forecast format as the between-participants variable. (Note that this analysis used high-error-first participants.) Participants rated high-error trials significantly lower ($M = 2.02$, $SD = .91$) than low error trials ($M = 2.36$, $SD = 1.09$), $F(1,$

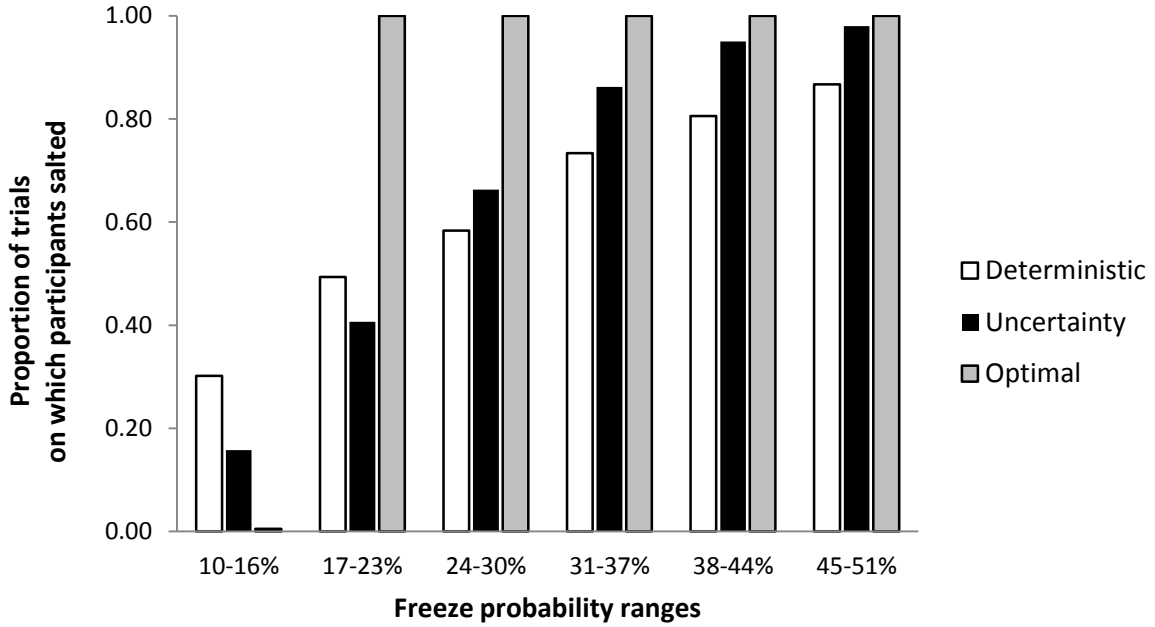
130) = 12.89, $p < .01$. Neither the main effect of format nor the interaction reached significance. Participants gave significantly lower trust ratings in the low-error trials, $t(282) = 4.34$, $p < .01$, when the low-error trials were presented after the high-error trials ($M = 2.36$, $SD = 1.09$) than when they were presented first ($M = 2.86$, $SD = .86$).

Uncertainty participants were more successful than deterministic participants at appropriately discriminating between when salting was warranted and when it was not. This was determined by analyzing a subset of low-error trials around the cost-loss threshold of 17% probability of freezing, from 10% to 23%. An equal number of trials in this subset, 18, was below 17%, therefore not warranting precautionary action, and above 17%, therefore warranting precautionary action. Choosing to salt below the 17% threshold represented a risk-averse error (salting unnecessarily), and not salting above the 17% threshold represented a risk-seeking error. A mixed-model ANOVA on the proportion of these types of errors was performed, with error type (risk seeking and risk averse) as the within-participants independent variable and forecast format (freeze probability, freeze frequency, advice, and control) as the between-participants variable. The analysis revealed a significant main effect for error type, $F(1, 146) = 113.41$, $p < .01$, suggesting that participants overall made more risk-seeking than risk-averse errors. There was not a significant forecast format main effect, but the interaction was significant, $F(3, 146) = 5.06$, $p < .01$, suggesting that uncertainty participants made fewer risk-averse but more risk-seeking errors than deterministic participants did. See Table 2.4.

Participants with uncertainty estimates made salting decisions that were overall more optimal than those made by participants with single-value forecasts or decision advice. Following the normative decision rule, a decision maker would apply salt when the freeze probability was 17% or greater and withhold salt when in was less than 17%. On average,

uncertainty participants (freeze probability and freeze frequency combined) better differentiated when to salt versus withhold salt than did deterministic participants (advice and control participants combined). A mixed-model ANOVA on mean proportion of "salt" responses per participant, with freeze probability range (above and below 17%) as the within-participants variable and forecast format (uncertainty and deterministic) as the between-participants variable, revealed a significant main effect for freeze probability range, $F(1, 148) = 1270.00$, $p < .01$. Participants salted more above the 17% threshold ($M = .74$, $SD = .17$) than below it ($M = .21$, $SD = .20$). There was also a significant interaction, $F(1, 148) = 34.09$, $p < .01$, suggesting greater differentiation in salt decisions among those with uncertainty estimates. Uncertainty participants salted less often below the 17% threshold ($M = .16$, $SD = .18$) than did those with deterministic forecasts ($M = .26$, $SD = .20$), and more often above it ($M = .77$, $SD = .14$) than did those with deterministic forecasts ($M = .70$, $SD = .18$). See Figure 2.2.

Figure 2.2. Decisions to apply salt over freeze probability ranges by condition, including optimal decisions.



Additionally, inclusion of uncertainty estimates minimized the negative influence of forecast error and directly combated the bias toward risk seeking when forecast error increased. This was determined by comparing decisions in low- and high-error trials directly. Although the sequences of low-error trials and high-error trials had the same set of deterministic forecasts, ranging from 32°F to 35°F, the probability of freezing in the high-error trials was somewhat higher ($M$ = 38%) than in the matching low-error trials ($M$ = 27%). For that reason, I analyzed a subset of 22 trials in which the appropriate course of action, to salt (PoF ≥ 17%), was the same in both the low- and high-error trials. A mixed-model ANOVA on the proportion of participants salting, with forecast format (uncertainty and deterministic) as the between-participants variable and forecast error (low and high) as the within-participants variable, revealed a significant main effect for forecast format, $F(1, 148) = 4.93$, $p = .03$. Participants with uncertainty forecasts salted more often ($M$ = .68, $SD$ = .19) than did participants with deterministic forecasts ($M$ = .61, $SD$ = .20). Moreover, there was a significant forecast error by format interaction, $F(1, 148) = 21.11$, $p < .01$. Participants with uncertainty formats salted more often in the high-error trials ($M$ = .72, $SD$ = .24) than they did in the low-error trials ($M$ = .64, $SD$ = .18), while participants with deterministic formats salted *less* often in the high-error trials ($M$ = .58, $SD$ = .24) than they did in the low-error trials ($M$ = .63, $SD$ = .18). The decrease in decisions to salt on high-error trials among deterministic participants suggested an increase in risk seeking. Uncertainty participants, however, took action more often despite the increase in forecast error.

Finally, participants' responses to the follow-up question about strategy were evaluated. (Decision advice participants were excluded from this analysis.) Participants responded to the following question: "What strategy did you use for making your decision to salt or not salt the roads?" Of particular interest was if participants used thresholds to make their decisions and if

their decisions were based on the cost-loss ratio. As such, responses were coded as "used likelihood threshold," "used temperature threshold," "used cost-loss ratio," or "other." (To be coded as "used cost-loss ratio," participants' responses needed to mention the basic idea of calculating a threshold by dividing the cost of salting by the potential penalty. They did not need to use the term "cost-loss ratio," nor did they need to correctly calculate a value of 1/6.) "Other" included a wide range of strategies, such as trying to identify a pattern from one trial to the next and guessing at random, plus blank responses and responses that were not possible to interpret. See Tables 2.6 and 2.7 below for frequency of strategies used and mean thresholds reported.

Table 2.6. Frequency of strategies used by forecast format.

| **Forecast format** | **Used likelihood threshold*** | **Used temperature threshold** | **Used both probability and temperature thresholds** | **Used cost-loss ratio** | **Other** | **Total** |
|---|---|---|---|---|---|---|
| Freeze probability | 16 | 2 | 4 | 0 | 55 | 77 |
| Freeze frequency | 10 | 6 | 4 | 2 | 56 | 78 |
| Control | na | 21 | na | na | 50 | 71 |

*For freeze frequency participants, "likelihood threshold" refers to the frequency expression (number of days out of 100); for freeze probability participants, it refers to the percent expression (percent of days).

Table 2.7. Mean reported threshold by forecast format.

| **Forecast format** | **Mean temperature threshold (SD)** | **Mean frequency or percent threshold (SD)** |
|---|---|---|
| Freeze probability | 34.00° (.89) | 30.60% (8.12) |
| Freeze frequency | 33.40° (.97) | 26.56 days out of 100 (10.23) |
| Control | 34.14° (1.32) | na |

Coding revealed that the number of participants employing likelihood, temperature, or cost-loss strategies was too low to allow for meaningful inferential statistical analysis. However, the frequency table reveals two interesting things. First, it simply shows that when available,

participants used the uncertainty estimates. The proportion of participants using just the temperature information to make their decisions was much lower in the uncertainty conditions than in the control condition. Moreover, it shows that even when given explicit uncertainty estimates, participants very rarely compared the estimates to the cost-loss ratio in order to make salting decisions, suggesting that participants did not calculate a cost-loss ratio or did not understand the underlying concept of weighing costs and losses to determine an optimal long-run threshold. Participants' decision thresholds also reveal important aspects of participants' understanding of weather uncertainty. First, the fact that there was variance in participants' reported decision thresholds suggests that participants had different risk tolerances. Second, evaluation of participants' stated decision thresholds revealed that the modal temperature threshold was 34°F, two degrees above the freeze threshold, suggesting that participants recognized the uncertainty inherent in the forecasts. If participants had not intuited weather uncertainty, one would have expected the modal (and mean) temperature decision threshold to be 32°F, the freezing point. Instead, many participants intuited the possibility of forecast error and therefore regarded a forecast of 34°F as close enough to freezing to take precautionary action.

A number of caveats must be noted. Participants' self-report of strategies used may or may not reflect the strategies they actually used. Participants might have used strategies that they did not report, including the strategies of present focus, like using a temperature threshold. Similarly, a participant who indicated basing his or her decision on a freeze probability estimate might have also used a temperature estimate without reporting doing so, and it is possible that the participant did not employ the same strategy on all trials. Furthermore, participants' responses were often difficult to interpret or code. For example, some participants reported the highest temperature at which they would apply salt, whereas others reported the lowest

temperature at which they would withhold salt. Many participants indicated that the boundary between salting and not salting was not a clear-cut threshold, e.g., "If the overnight low for temperature was over 35 degrees I chose not to salt a majority of the time. If the temperature was 32 degrees I chose to salt every time. In between 32-35 I chose to salt 70% of the time." Participants rarely reported a distinct threshold, e.g., "Temp above 35 no salt, below 35 salt," and even then, it remained unclear what those participants would do if the forecast were at the threshold, e.g., 35°. Therefore, the above data are just a rough indication of strategy.

**Discussion**

The results revealed three main things: 1) participants who were given forecasts that included estimates of uncertainty made better decisions and trusted the forecasts more than participants who were given only single-value forecasts; 2) increased forecast error led to a deterioration in participants' performance and trust, but that effect was attenuated for participants with uncertainty estimates; 3) providing explicit decision advice did not affect participants' performance: Participants' decisions were no better (or worse) than those of participants who were given single-value forecasts, and their ratings of trust were no different, either.

The superior decision making of uncertainty participants is the most fundamentally important result. It contributes to the growing body of empirical evidence that, contrary to much past research in cognition, uncertainty estimates can benefit decision makers, and, equally as important, decision makers have greater trust in forecasts that include uncertainty estimates than in forecasts that do not. This is a significant replication of a critical empirical finding in applied psychology, one which carries practical implications for how weather-related risk is communicated.

That inclusion of uncertainty estimates attenuated performance deterioration caused by forecast error is a compelling justification for their inclusion in forecasts. While participants in all conditions made inferior decisions when the forecast error was doubled, participants with uncertainty estimates suffered significantly less than those with only single-value forecasts. Participants with single-value forecasts tended toward inaction when forecast error increased, consistent with past research (Erev & Barron, 2005). This result strongly suggests that part of the benefit of the inclusion of uncertainty estimates is that such forecasts seem less inaccurate than single-value forecasts when the forecasts fail to verify, i.e., when the forecasts and observations do not match. Single-value deterministic forecasts, on the other hand, are substantially less trustworthy. The very idea of determinism – that a predicted event *will* happen – is significantly more false when the forecasts and observations are badly mismatched. Participants' responses to the open-ended question about strategy suggested that they inferred that there was uncertainty around the single-value forecasts, and as such, single values were regarded as insufficient to make good decisions (or decisions at the level of those made by participants with uncertainty estimates), and trust suffered as a consequence.

Additionally, at least for the types of weather-related decisions made in the road salting experiment, uncertainty representation did not matter. Overall, freeze probability participants and freeze frequency participants demonstrated similar decision quality and trust. It might be that the sort of difficulty with probability that people have demonstrated in past research, namely probabilities about single events or events that have never happened (Brase, 2002; Lipkus, 2007; Hoffrage et al., 2002), was not an issue with weather-related probability, as people might have intuition about weather records, an intuitive frequentist sense that there have been numerous past forecasts in similar conditions, allowing for a probability to be calculated. For instance, if the

forecast for tonight's low temperature is 37°F, people probably know that there have been thousands of such forecasts of 37°F over time for a given area, and on a certain proportion of those, a freezing temperature was observed. As such, in this situation, perhaps stating "a 25% chance of freezing" is no more difficult to understand than "it will freeze 25 out of 100 days like this." It might be that this is not a problem in which representation matters.

The results demonstrated that risk seeking was the dominant type of decision error, as predicted. Freezing temperatures were rarely observed, which likely led to participants underweighting their likelihood. The framing of the task might have exacerbated the risk-seeking bias, as well. Participants chose between a sure cost or the chance of a greater loss, and they tended, overall, to take the gamble. Although the inclusion of uncertainty estimates attenuated risk-seeking errors when forecast error increased, and although uncertainty participants were better than deterministic participants at discriminating between when salting was warranted and when it was not, uncertainty participants were not less risk-seeking overall. There was not a significant main effect for forecast format on errors; there was only a significant interaction between forecast format and error type, which showed that uncertainty participants made fewer risk-averse errors. Risk seeking among uncertainty participants was particularly notable at low probabilities of freezing for which precautionary action was warranted. Figure 2.2 reveals that in the range of 17 to 23% probability of freezing, uncertainty participants actually took less precautionary action than deterministic participants did. Thus, in the low range of freeze probability for which precautionary action was necessary, control and advice participants actually did better than uncertainty participants.

However, providing participants with economically optimal advice – a simple statement of binary choice, salt or no salt – did not improve performance at all above not providing advice.

The decision advice manipulation was intended to capture the real-life situation of emergency managers or other authorities giving explicit advice (e.g., evacuation orders) to people threatened by extreme weather. As field research shows that may people do not heed official weather warnings, perhaps it is not surprising that most advice participants did not comply with the DSA's recommendations, but it does seem surprising that there was *no* measurable improvement over not having advice at all. This issue was explored further in Experiment 2.

**Experiment 2: Elaboration of Decision Advice**

My finding that participants did no better at the road salting task when they were shown explicit, optimal decision advice than when they were shown a mere single-value forecast came as quite a surprise. Although the relatively poor performance among advice participants might indeed be analogous to the real-life phenomenon of the public's disregard of weather warnings, surely the inclusion of decision advice must confer *some* benefit to decision makers. However, the decision advice was based on a threshold – the cost-loss ratio threshold – that was never described in any way to the participants. Perhaps participants in Experiment 1 did not trust the decision advice because the advice effectively made assumptions about the participants' own risk tolerances and salting thresholds (Roulston & Smith, 2004). Additionally, because the normative decision threshold based on the cost-loss ratio was so low (17%), the DSA often – usually – advised salting, and most of the time salting turned out to be unnecessary. Therefore, it is reasonable to expect that if participants were given a mathematical explanation about the DSA's decision-generating process, they might heed its advice more. Alternatively, participants might prefer a more intuitively compelling narrative that emphasized not the mathematical calculation underlying the advice but instead the long-term benefit of following the advice, even if it did not lead to optimal decisions from one trial to the next.

In Experiment 2, the road salting task included two new manipulations. For some participants, task instructions included a mathematical explanation of the manner in which the DSA advice was generated; for other participants, the instructions included a narrative (i.e., non-mathematical) explanation of the benefit of following the advice. I hypothesized that one or both of these elaborations of the decision advice would lead to greater trust and compliance, resulting in better decisions and better performance at the task.

**Method**

Participants. One hundred sixty-nine participants (47.8% female) took part in the experiment. Age ranged from 18 to 29 years ($M$ = 19.1). All participants were recruited from introductory psychology courses at the University of Washington and were awarded a small amount of academic credit for participating. Participants signed up voluntarily and had a chance to win a small cash prize.

Apparatus. Exactly as before, the experiment was programmed with Microsoft Excel Visual Basic and was administered via Excel on standard desktop computers.

Procedure. The procedure in Experiment 2 was identical to that in Experiment 1.

Stimuli. The stimuli used in Experiment 2 were the same as those used in Experiment 1.

Design. This version of the experiment included a single between-participants independent variable, forecast format. There were four levels: advice, like in Experiment 1; advice with a mathematical explanation of how the DSA's decision recommendation was calculated; advice with a narrative explanation of the long-run benefit of following the DSA's recommendations; and a control, which included only a single-value forecast and no decision advice.

In the two new advice conditions, the forecasts were identical to the forecasts in the original advice condition: There was simply a message expressing the DSA's treatment recommendation. The manipulation occurred in the instructions, in which additional information was presented to participants. For the advice/mathematical explanation, it was an explanation of the cost-loss ratio and how that underlay the DSA's recommendations; for the advice/long-run narrative, it was an explanation of the benefit of following the DSA's advice in the long run, even if individual trials resulted in unnecessary (or missed) precautionary action. The same information was presented again on the "break" screens between each virtual month of the experiment. See Table 2.8.

Table 2.8. Mathematical explanation and long-run narrative included in participants' instructions for those conditions.

| **Forecast format** | **Text included in instructions beyond basic description of the decision advice** |
|---|---|
| Advice with mathematical explanation | Here is how DSA decides when to tell you to salt:<br><br>It compares the cost of salting to the penalty for not salting weighted by the probability that you will get penalized if you do not salt (the probability of freezing).<br><br>• Salting costs $1,000.<br>• Not salting costs nothing, unless a freezing temperature is observed, which results in a penalty of $6,000.<br>• Therefore, the cost of not salting is actually $6,000 x probability of freezing.<br>• Weighing the cost of salting against the cost of not salting, there is a break-even point in probability of freezing of<br>16.67% ($1,000 / $6,000 = .1667, or 16.67%).<br>• If the probability of freezing is less than 17%, you should not salt; if it's greater than or equal to 17%, you should salt.<br><br>The DSA accounts for probability of freezing and tells you:<br>Salt if probability of freezing ≥ 17%<br>Do not salt if probability of freezing < 17% |
| Advice with long-run narrative | This advice will help you maximize your budget over the long run. It is important to understand that this advice is probabilistic. That means that for individual days you may salt sometimes when it does not freeze. That is OK to do because the penalty for not salting is so large that it works out in the long |

| | |
|---|---|
| | run. In other words, the benefit of following the rule will not be seen every day but will be realized over the long run. This is similar to a fair coin toss. The probability of getting heads in tossing a coin is 50%, but this does not mean that if you toss a coin 10 times you will always get 5 heads and 5 tails. You could get 10 heads or 10 tails. However, if you continued to flip the coin 1,000 times, you would get approximately half heads and half tails.<br><br>The advice will be given in each trial. |

**Results**

It was predicted that supplementing the decision advice with explanatory information, whether in the form of a mathematical explanation of the decision rule or a narrative explanation of the long-run benefit of following the advice, would lead to improved decision quality and higher trust over participants with the basic decision advice and single-value forecast. Before analysis, five outlier participants were removed by the same data cleaning procedure employed in Experiment 1, leaving 164 participants for subsequent analysis.

Next, mean expected decision value and trust were analyzed. In Experiment 2, although high-error trials were included, the research question of interest did not concern differences in decision quality between low- and high-error trials, so the following analyses use only low-error trials, from both Month 1 and Month 2.

Adding explanations to the decision advice had no effect. Expected value scores were calculated and averaged in the same manner as they were in Experiment 1. Difference scores were not calculated because the forecast error variable was not explored, i.e., all trials were low-error trials. An ANOVA on mean expected values, with forecast format (advice with mathematical explanation, advice with long-run narrative, advice alone, and control) as the independent variable, revealed no significant effect, $F(3, 163) = .46$, $p$ = ns. See Table 2.9.

Again, adding explanations to the decision advice had no effect. Mean trust ratings were calculated from the end-of-month trust ratings for Month 1 and Month 2. An ANOVA on mean trust ratings revealed no significant effect, $F(3,163) = 1.95$, $p$ = ns. See Table 2.9.

Decision errors were analyzed by comparing error rates in trials with probabilities of freezing near the 17% cost-loss threshold. As before, a risk-averse error was salting when the probability of freezing was between 10% and 17%, and a risk-seeking error was withholding salt when the probability of freezing was between 17% and 23%. Participants made more risk-seeking errors than risk-averse errors. Interestingly, participants with decision advice made fewer errors than did control participants. A mixed-model ANOVA on proportion of trials on which participants made a decision error, with error type (risk averse and risk seeking) as the within-participants variable and forecast format (advice with mathematical explanation, advice with long-run explanation, advice alone, and control) as the between-participants variable, yielded a significant main effect for error type, $F(1, 160) = 58.56$, $p < .01$, suggesting participants made more risk-seeking than risk-averse errors. The significant effect for forecast format, $F(3, 160) = 4.51$, $p < .01$, suggested participants who received decision advice made fewer decision errors than participants who did not. Tukey's post hoc test showed that that effect was driven by significant differences between controls ($M = .41$, $SD = .11$) and both advice alone ($M = .35$, $SD = .08$), $p = .01$, and mathematical explanation participants ($M = .34$, $SD = .09$), $p < .01$. The interaction between forecast format and error type was not significant, although, importantly, a direct comparison of risk-seeking errors between advice alone and control participants suggested advice participants made significantly fewer risk-seeking errors than control participants, $t(70.17) = 2.12$, $p = .04$ (adjusted for unequal variances). See Table 2.9.

Table 2.9. Mean (standard deviation) expected decision value, mean trust ratings, and decision errors for Experiment 2.

| **Forecast format** | **Mean expected decision value** | **Mean month trust rating** | **Risk-averse error rate** | **Risk-seeking error rate** |
|---|---|---|---|---|
| Advice with mathematical | -$1,074.77 ($75.56) | 2.41 (.76) | .20 (.18) | .48 (.23) |
| Advice with long-run | -$1,074.34 ($89.05) | 2.37 (.88) | .24 (.22) | .50 (.26) |
| Advice alone | -$1,081.95 ($80.48) | 2.40 (.76) | .27 (.18) | .43 (.17) |
| Control | -$1,094.27 ($100.11) | 2.05 (.66) | .29 (.21) | .52 (.23) |
| *mean* | *-$1,081.18 ($86.19)* | *2.31 (.78)* | *.25 (.20)* | *.48 (.23)* |

**Discussion**

Elaborating the decision advice with a mathematical explanation or a compelling long-run-benefits explanation neither increased trust nor improved decision quality. Participants still generally ignored the advice and performed no differently from participants who did not have decision support.

That additional information about the DSA did not help suggests that it is not just a matter of users wanting more information. Rather, it might be that participants are better served by better information upon which personally relevant decisions can be based. Comparing the forecast formats tested in Experiments 1 and 2, uncertainty estimates appear to constitute better information than decision advice. In Experiment 1, participants with uncertainty estimates – better information that participants could use to make decisions that fit their own risk tolerances – had higher trust in the forecasts and performed the task better than participants without that information. In Experiment 2, participants were given decision advice that was based on a fixed, externally defined risk tolerance, in this case on the cost-loss ratio threshold, which likely

seemed less personally relevant to the decision makers. Responses to the strategy question revealed that a variety of decision thresholds was used, demonstrating both that many participants perceived uncertainty in the forecasts and that participants' individual risk tolerances differed. Both of these factors were directly served by inclusion of uncertainty estimates. Participants who were shown uncertainty estimates had better information to use in accordance with their own risk tolerances, and the uncertainty estimates were consistent with participants' intuitions about weather uncertainty, thereby generating trust in the forecasts.

However, in Experiment 2, a benefit of decision advice emerged: Participants who were given decision advice made fewer decision errors than participants without advice. Advice participants, including participants who were given the advice alone without explanation, made fewer risk-seeking errors than control participants did. In Experiment 1, inclusion of uncertainty estimates was beneficial in all ways (e.g., decision quality, trust) *except* in reducing risk-seeking decision errors. Adding explicit decision advice might therefore be helpful in prompting action on trials with low freeze probabilities for which salting is warranted. Might people benefit most from forecasts that include both uncertainty estimates and decision advice? This question was explored in Experiment 3.

**Experiment 3: Decision Advice with Uncertainty Estimates**

Participants did not perform the task better or give higher ratings of trust when given explicit decision advice, even when the advice was explained either in terms of the mathematics of the cost-loss ratio or in terms of the long-run benefit. However, participants with decision advice did make fewer decision errors overall, including risk-seeking errors. While uncertainty participants in Experiment 1 demonstrated better task performance and higher trust, they made

more risk-seeking errors, suggesting that including advice might reduce risk-seeking errors and prompt participants to take precautionary action at low probabilities for which such action is warranted.

To explore the effect of combining decision advice and uncertainty estimates, I ran the road salting experiment again, this time with a condition in which decision advice and uncertainty estimates were combined.

**Method**

Participants. One hundred seventy-eight participants (48.3% female) took part in the experiment. Age ranged from 18 to 28 years ($M$ = 19.5). All participants were recruited from introductory psychology courses at the University of Washington and were awarded a small amount of academic credit for participating. Participants signed up voluntarily and had a chance to win a small cash prize.

Apparatus. Exactly as before, the experiment was programmed with Microsoft Excel Visual Basic and was administered via Excel on standard desktop computers.

Procedure. The procedure in Experiment 3 was identical to that in the previous experiments.

Stimuli. The stimuli used in Experiment 3 were the same as those used in the previous experiments.

Design. This version of the experiment included a single between-participants independent variable, forecast format. There were three levels: advice plus freeze probability, advice plus freeze frequency, and a control, which included only a single-value forecast and no decision advice. See Table 2.10 with forecast expressions.

Table 2.10. Example forecasts for each forecast format.

| **Forecast format** | **Example forecast** |
|---|---|
| Advice + freeze probability | "The expected nighttime low temperature is 35°F; there is a 20% chance the temperature will be 32°F or less. DSA: 'Salting is recommended under these circumstances.'" |
| Advice + freeze frequency | "The expected nighttime low temperature is 35°F; 20 out of 100 days like this, the temperature will be 32°F or less. DSA: 'Salting is recommended under these circumstances.'" |
| Control | "The expected nighttime low temperature is 35°F." |

**Results**

It was hypothesized that adding uncertainty estimates to the decision advice would result in the highest decision quality and trust ratings yet tested. First, eight outlier participants were removed from subsequent analysis by the means described before, leaving 170 participants for analysis. Like with Experiment 2, analysis for Experiment 3 concerned only low-error trials (Month 1 and Month 2).

Adding uncertainty estimates to the decision advice significantly improved decision quality and trust. An ANOVA on mean expected value, with forecast format as the independent variable, revealed a significant effect, $F(2, 169) = 15.72$, $p < .01$. Tukey's post hoc analysis showed that advice plus freeze probability and advice plus freeze frequency participants demonstrated decision quality significantly superior to that of control participants, $p < .01$ in both cases. An ANOVA on mean end-of-month trust rating also yielded a significant effect, $F(2, 169) = 3.42$, $p = .04$. Tukey's post hoc test revealed that advice plus freeze frequency participants trusted the forecasts significantly more than control participants did, $p = .05$, but there was no statistically significant difference between advice plus freeze probability participants' and control participants' trust ratings. See Table 2.11.

Table 2.11. Mean expected decision value and mean trust ratings for Experiment 3.

| **Forecast format** | **Mean expected decision value** | **Mean month trust rating** | **Risk-averse error rate** | **Risk-seeking error rate** |
|---|---|---|---|---|
| Advice + freeze probability | -$1,033.85 ($72.48) | 2.47 (.78) | .17 (.19) | .60 (.23) |
| Advice + freeze frequency | -$1,053.65 ($79.25) | 2.77 (.75) | .13 (16) | .51 (.25) |
| Control | -$1,121.09 ($104.65) | 2.43 (.72) | .27 (.19) | .55 (.21) |
| *mean* | *-$1,069.86 ($93.93)* | *2.56 (.76)* | *.19 (.19)* | *.56 (.23)* |

Finally, decision errors were analyzed. Like before, participants made more risk-seeking than risk-averse errors. Advice combined with uncertainty estimates led to fewer errors overall, and, importantly, to a rate of risk-seeking errors that was no greater than that of controls. A mixed-model ANOVA on the proportion of trials on which participants made errors, with error type as the within-participants variable and forecast format as the between-participants variable, revealed a significant main effect for error type, $F(1, 167) = 160.69$, $p < .01$. Participants made more risk-seeking than risk-averse errors. There was a significant main effect for format, $F(2, 167) = 8.86$, $p < .01$. Tukey's post hoc analysis indicated that there were significantly smaller proportions of errors in both the advice plus freeze frequency ($M = .37$, $SD = .09$, $p = .02$) and advice plus freeze probability ($M = .34$, $SD = .10$, $p < .01$) conditions than in the control condition ($M = .41$, $SD = .08$). There was also a significant interaction, $F(2, 167) = 4.22$, $p = .02$, suggesting that the error type effect was driven primarily by low risk-averse error rates among combined-formats participants; risk-seeking errors were similar across conditions. See Table 2.11.

## Discussion

Experiment 3 provides compelling evidence of the value of combining uncertainty estimates with decision support. Combining the formats preserved both the principal advantage of the forecast with uncertainty estimates (better overall task performance and higher trust) and the decision advice (less risk-seeking at low probabilities for which precautionary action was merited). In Experiment 1, uncertainty participants did better than control participants but did not make fewer decision errors overall compared to control participants. In Experiment 3, however, the combined-format participants did better than controls *and* made fewer overall decision errors.

In Experiment 1, participants with uncertainty estimates might have regarded spending money on precautionary action as unnecessary at low freeze probabilities for which precautionary action was warranted, between 17% and 23%. Unless one understood the concept of expected value and calculated the appropriate cost-loss ratio, which Experiment 1's strategy data suggested was largely not the case, a probability just above 17% might have seemed too low to warrant precautionary action. Results from Experiment 1 suggest that participants with uncertainty estimates did best overall, but the risk of an overnight freezing temperature at low probabilities for which precautionary action was needed was better communicated with the addition of explicit decision advice, as demonstrated in Experiment 3.

By itself, the Decision Support Aid was not an effective means to guide decisions, and this might in part be what is behind people's reluctance to take precautionary action in real-life situations of weather-related risk. People might not trust weather warnings, perhaps because people have different risk tolerances from one another and as such they might simply regard the weather warnings as applying to (or being generated from) other risk tolerances. Furthermore,

these “black-box” decision recommendations might seem untrustworthy, as individual decision makers are not told explicitly what criteria distinguish particular warnings (e.g., evacuate vs. do not evacuate). However, even providing that information, whether in the form of an analytical explanation or a more intuitively compelling narrative, might not improve compliance with weather warnings. Such additional information did not improve decision performance or engender more trust among participants in the experimental setting.

## General Discussion

These three experiments produced clear evidence of the benefits of including numeric uncertainty estimates in weather forecasts. Additionally, I was able to demonstrate some patterns of behavior observed in numerous field studies, such as not heeding explicit decision advice, in a controlled laboratory setting.

The studies offered strong evidence that supports the growing literature about people’s ability to make practical and effective use of uncertainty estimates. Experiment 1 showed how including uncertainty estimates improved decision quality overall and lessened the deterioration of decision quality when forecast error increased, as compared to single-value forecasts. Perhaps most importantly, the results showed that including uncertainty estimates led to higher ratings of trust in the forecasts, perhaps by making forecast-observation mismatches seem less inaccurate, and perhaps also because the inclusion of uncertainty estimates matched users’ intuition about weather uncertainty, thereby making the forecasts seem more complete and the forecasters more forthright. This is critically important because of the potential it offers forecasters, who are concerned that forecast users will begin to distrust forecasts if forecasted events do not occur (NRC, 2006). Including uncertainty estimates acts as a sort of insulation against distrust, even

when the forecasts are of lesser quality. This is especially important in situations in which there is often high forecast error: early warnings about developing or approaching storms, in which forecast information must be communicated with sufficient lead time to allow vulnerable people to take precautionary action. In Experiment 1, when the high-error block of trials was presented before the low-error block, it was demonstrated that once trust was lost, it could not be regained, consistent with the claims of others (Slovic, 1999), so preserving forecast users' trust ought to be a top priority of forecasters and emergency managers when dangerous weather situations arise.

Experiment 1 also provided a clear demonstration of people's tendency to take risks when faced with potential losses. The road salting task features a cost structure framed entirely in terms of losses: There is a cost to take precautionary action, and if that cost is not taken, there is a potential loss. This is the same situation in most real-life weather-related situations. As such, my controlled laboratory experiment demonstrated what often happens in real life: When faced with a cost-loss situation, people have a tendency to take a gamble, to seek more risk than is warranted. Overall, providing experimental participants with uncertainty estimates led to more economically optimal decisions, which carries implications about how decision makers in real-life situations might perceive weather-related risk differently if they were presented with uncertainty estimates.

Finally, the experiments clearly showed the lack of effectiveness of explicit decision advice to help with a weather-related task, even when the mechanism underlying the system was made transparent, *except* when used to prompt precautionary action at low probabilities. Combining the decision advice with numeric uncertainty estimates led to the highest decision quality and lowest decision error rate. The benefit of combining the two might have been due to people's preference to exercise their own risk tolerances while still having normative advice to

include in the decision process. On its own, the decision advice was neither a trusted nor a helpful information source. The following chapter presents an experimental test of what might have been at least partly responsible for the relatively poor performance of advice participants: the effect of false alarms.

# Chapter 3: The False Alarm Effect

## Background

The results from the advice manipulation in the decision advice and forecast error experiments were particularly troubling: Directly informing participants how to decide when faced with uncertainty led to no meaningful improvement in task performance over leaving participants on their own with only the most basic forecast information. Even explaining to them how the decision advice was generated and that there was a long-run benefit of following the advice did not make any difference. Participants ignored the decision advice. Indeed, it is possible that this laboratory finding is simply reproducing what occurs in real life: Many people do not heed warnings of approaching severe weather, even when advised to take an explicit course of action (e.g., evacuate).

One theoretical explanation for why people fail to heed weather warnings is the false alarm effect, or the cry wolf effect (Breznitz, 1985), named after Aesop's famous fable, "The Boy Who Cried 'Wolf'":

> *There was a boy tending the sheep who would continually go up to the embankment and shout, "Help, there's a wolf!" The farmers would all come running only to find out that what the boy said was not true. Then one day there really was a wolf but when the boy shouted, they didn't believe him and no one came to his aid. The whole flock was eaten by the wolf. The story shows that this is how liars are rewarded: even if they tell the truth, no one believes them.* (Gibbs, 2002)

Similarly, across any number of situations or domains, when people are repeatedly warned about something that does not come to pass, they might begin to discount the warnings (Dow & Cutter, 1998; Whitehead et al., 2000). This is potentially an extremely dangerous problem in the case of weather warnings, as the events being warned about are often lethal. There is significant concern

that forecasts will be come to be distrusted if forecasted events do not occur (NRC, 2006). Furthermore, false alarms are a relatively common type of forecast error: Forecasters tend to "overforecast" in order to avoid misses (not warning of an event that ultimately does occur), as misses are generally perceived to be more serious and costly than false alarms (Murphy, 1991).

Much past research suggests a negative effect of receiving false alarms. False alarms are said to lead users of warning systems to "disuse" or ignore warning systems, rendering the systems largely useless (Parasuraman & Riley, 1997). Consistent with this, one empirical study found that the lower the predictive value of an alarm (i.e., the greater the false alarm rate), the slower were users' responses to the alarm (Getty et al., 1995), a result similar to those found in a study of maritime students tasked with using an automated system to identify disturbances (Kerstholt & Passenier, 2000) and in a medical diagnosis study (Kerstholt, 1995). Research on warning systems for car drivers suggested that alarm systems with lower sensitivity (and thus fewer false alarms) led to the fewest adverse driving events (Gupta, Bisantz, & Singh, 2001) and that compliance with such warning systems was greater for systems with fewer false alarms (Cotté, Meyer, & Coughlin, 2001). False alarms have been demonstrated to affect decision making in several ways, e.g., delayed action (Kerstholt, 1995; Kerstholt & Passenier, 2000; Getty et al., 1995), taking all-or-nothing approaches to action (Bliss, Gilson, & Deaton, 1995), and using probability matching (Bliss & Dunn, 2000).

Importantly, research from the field often finds that when faced with the threat of severe weather, like hurricanes, the cry wolf effect often figures into residents' decision not to evacuate, even when they are given official evacuation notices (Burnside, 2006). However, some note that prospective evacuees prefer to make decisions based more on their individual risk tolerances,

i.e., it is not so much that they are false alarm intolerant as they do not feel that official warnings take their personal situations into account (Dow & Cutter, 1998).

However, some research suggests that in certain situations, users can tolerate a certain amount of false alarm. Analysis of field data from hurricane evacuees suggests that the majority of those who evacuate unnecessarily will evacuate again in the face of future emergencies (Baker, 1991). Furthermore, the ability to recognize the cause of a false alarm has been associated with false alarm tolerance (Cotté, Meyer, & Coughlin, 2001). In summary, it is not entirely clear from past research how false alarms affect weather-related decisions. Past findings, taken mostly from the field and often from domains other than weather, have been mixed.

Within the context of the previous experiments reported here, it is readily believable that the failure of the decision support aid was a demonstration of a false alarm effect. Because the cost-loss ratio was quite low (17%), the DSA was usually advising applying salt treatment, but freezing temperatures were usually not observed. In the 60 low-error trials of the previous experiment, the DSA recommended salting 70% of the time (42 times), but a freezing temperature was observed on only about 40% of those trials (17 trials). In other words, almost 60% of the DSA's salting recommendations were false alarms (25 out of 42 trials). Participants likely perceived that they were wasting their money by following the DSA's advise; indeed, on average, participants followed the advice for only three trials before abandoning it.

It has been suggested, however, that weather-related false alarm effects can be reduced by raising the threshold at which weather warnings are given, thereby lowering the false alarm rate (Roulston & Smith, 2004). For example, if an emergency manager had a set forecast probability threshold of 20% (for whatever reason), meaning that he or she would issue a

warning if the probability of the event exceeded 20%, then he or she could reduce the number of false alarms generated over time by raising the warning threshold to 25%. The resultant increase in misses would theoretically be offset by an increase in users' trust and consequent compliance with the warnings. This strategy to increase compliance with weather warnings has never before, to my knowledge, been tested in a controlled laboratory setting.

An alternative strategy to adjusting the false alarm rate is simply to augment decision advice with a probabilistic estimate of uncertainty. The previous experiments demonstrated that inclusion of an uncertainty estimate resulted in an increase in user trust and in higher decision quality. Perhaps forecast users would be better served if they were presented with accurate, usable information (i.e., uncertainty estimates) than if they were presented with arguably misleading information (i.e., decision advice based on non-optimal warning thresholds). Furthermore, there is evidence from field research that a combination of decision advice and uncertainty estimates would be especially beneficial to non-expert decision makers. There is evidence both for the effectiveness of official warnings (e.g., Baker, 1991; Baker, 1995) and for the importance of personal relevance (e.g., Dow & Cutter, 1998) in the extreme-weather-related decision-making process. Decision advice would provide decision makers with categorical directive, and uncertainty estimates would cater to decision makers' diverse risk tolerances.

To compare the false-alarm-level and uncertainty-estimate strategies to improve decision quality and compliance with weather warnings, I performed an experiment using the road salting task. Participants, as road maintenance company managers, had to choose whether to apply or withhold salt treatment based on weather forecasts over a series of trials. All participants received decision advice from a support aid, but the support aid's advice criterion was systematically manipulated to produce different rates of observed false alarms. For some

participants, the support aid advised applying treatment at very low probabilities of a freezing temperature, thereby generating a high rate of observed false alarms; whereas for other participants, the support aid advised treatment at increasingly high probabilities, generating lower observed false alarm rates. Additionally, some participants were also given probabilistic uncertainty estimates, whereas others were not. Decision quality, trust in the forecasts, and compliance with the support aid were measured.

## Method

To test the false alarm hypothesis, a new experiment was run using the same basic road salting task. However, the task was modified, both superficially and substantively, from the previous experiment. Superficial changes included rewording parts of the instructions for clarity; reducing the number of trials from 120 to 60, because forecast error was not manipulated in the present experiment and only one block of 60 forecasts was needed; changing the cash payout structure, described below; increasing from five to six the number of response options in the trust measures; and changing the cost structure: The cost to apply salt remained $1,000 but the penalty for not salting when a freezing temperature was observed was changed to $5,000, resulting in a cost-loss ratio of 20%. More substantively, a new forecast set was created, described below in Stimuli.

Participants. Three hundred eighty-eight participants (54.9% female) took part in the experiment. Age ranged from 18 to 36 years ($M$ = 19.3). All participants were recruited from introductory psychology courses at the University of Washington and were awarded a small amount of academic credit for participating and the chance to win a small cash prize. Participants signed up voluntarily.

Apparatus. The experiment was programmed with Microsoft Excel Visual Basic and was administered via Excel on standard desktop computers.

Procedure. Participants arrived in groups of 1 to 12 at pre-arranged times in a computer lab. After participants read a consent form and agreed to participate, the experimenter presented task instructions and a practice trial while participants followed along on their computers. Following an opportunity for questions, the task began. Participants completed the task individually. Upon completion of the task, participants answered some follow-up questions and were dismissed. All participants who showed up for the experiment were awarded credit.

Stimuli. The forecast set used in the false alarm experiment was similar to that used in earlier iterations of the road salting task. Real forecasts from Spokane and Yakima were used again, observed temperatures were calculated the same way that had been before, and realistic trial sequences were again created. The primary difference was the distribution of temperatures and probabilities of freezing, a change which was necessary to have sufficient variance at the different false alarm levels. See Table 3.1.

Table 3.1. Forecast characteristics.

| | **Forecasts in original road salting experiment (low-error trials)** | **Forecasts in false alarm experiment** |
|---|---|---|
| **Mean forecast error** | 3.26°F | 3.17°F |
| **Mean freeze probability** | 24.75% (range 10-51%) | 29.00% (range 8-51%) |
| **Mean forecast** | 34°F (range 32-37°F) | 34°F (range 32-37°F) |
| **Mean observation** | 35°F (range 27-42°F) | 34°F (range26-41°F) |

Design. The false alarm experiment had an incomplete factorial design. There were two between-participants independent variables: forecast format and false alarm level. Forecast format had two levels: advice, in which participants were presented with overnight low temperature forecasts and a treatment recommendation (apply salt or withhold salt) from the Decision Support Aid; and advice plus freeze probability, in which the forecast and treatment advice were augmented with an estimate of uncertainty, expressed as the probability of a freezing temperature. False alarm level had four levels: High FA-10, Unadjusted FA-20, Low FA-30, and Low FA-40, described below. The forecast format and false alarm level variables were fully crossed, resulting in eight experimental conditions. Additionally, there was a control, in which participants were presented only with the single-value forecast, resulting in a total of nine conditions.

In the present experiment, a false alarm was defined as an instance in which the Decision Support Aid recommended applying salt treatment but a freezing temperature was ultimately not observed. To vary the false alarm level, or the rate of non-freezing observations relative to DSA-generated treatment recommendations, I systematically raised or lowered the probability of freezing at which the DSA would advise treatment. For example, at the High FA-10 false alarm level, I programmed the DSA to issue salt treatment recommendations whenever the probability of freezing on a given trial was 10% or higher. That resulted in the DSA advising salt treatment on almost every trial, thereby generating a high observed false alarm rate. On the other end of the spectrum, at the Low FA-40 level, the DSA recommended applying salt treatment only when the freeze probability was 40% or greater, resulting in treatment being advised on relatively few trials and generating a low observed false alarm rate. With the cost-loss ratio of 20%, the economically optimal strategy would be to apply salt treatment whenever the freeze probability

was 20% or greater. At the Unadjusted FA-20 level, the DSA was programmed to do this. See Table 3.2.

Table 3.2. False alarm rate at each level of the false alarm level variable.

| **Condition: False alarm level** | **number of trials on which freezing temp observed** | **number of trials DSA advised treatment** | **number of trials DSA advised treatment and observation >32°F** | **False alarm rate** |
|---|---|---|---|---|
| High FA-10 | 18 | 56 | 38 | .68 |
| Unadjusted FA-20 | 18 | 45 | 29 | .64 |
| Low FA-30 | 18 | 30 | 18 | .60 |
| Low FA-40 | 18 | 15 | 8 | .53 |

The cash payout was modified for the false alarm experiment. Strictly following the advice given by the DSA would result in greater or worse performance depending on the false alarm level, with the Unadjusted FA-20 advice appropriately leading to the best performance and the Low FA-40 advice leading to the worst. Within the virtual world of the road salting task, the relative goodness or badness of following the given decision advice was reflected in the resulting balances. However, in the real world of people sitting in the computer lab and completing my experiment, in order to be fair to participants regardless of which condition they were in, participants were paid relative to the optimal final balance associated with following the advice they were given, rather than at a uniform, across-levels threshold. For example, if they followed the advice, Unadjusted FA-20 participants would end up with a relatively high balance, in a sense making the task easier for those participants than for Low FA-40 participants, who would end up with a relatively low balance if they followed the advice given to them. To eliminate this disparity, participants were paid cash for ending with condition-specific "optimal balances" (the balances associated with following the given advice). All participants had the same starting

balance, and the between-levels differences were adjusted for the threshold at which participants received payment. For example, at the High FA-10 level, the optimal ending balance was $14,000, so participants with balances of at least that amount received payment; at the Low FA-40 level, the optimal ending balance was $0, so participants with balances of at least that amount received payment. See Table 3.3. Participants were paid $3 for finishing the task with the optimal balance and earned an additional $1 for each additional $5,000 in their balance.

Table 3.3. Optimal final balances and mean expected decision values.

| **Condition: False alarm level** | **Optimal final balance** | **Optimal mean expected value** |
|---|---|---|
| High FA-10 | $14,000 | -$962.50 |
| Unadjusted FA-20 | $15,000 | -$920.00 |
| Low FA-30 | $10,000 | -$962.50 |
| Low FA-40 | $0 | -$1,137.50 |
| Control (single-value, no advice)* | $10,000 | -$1,175.83 |

*For the control condition, the optimal final balance was not "optimal," per se, but rather was set as the threshold for cash payment because that was the balance a participant would finish with if he or she applied salt treatment on every trial. Optimal mean expected value for control participants was based on choosing to apply treatment if the single-value forecast was 32°F or less and withholding it otherwise.

**Results**

A classic false alarm effect was predicted: As false alarm level increased, I hypothesized that decision quality, trust, and compliance would decrease. I also hypothesized that participants with decision advice and uncertainty estimates would make better decisions, have higher trust, and be more compliant than participants with advice alone. Before exploring these hypotheses, as with the other road salting experiments, I omitted outlier participants based on their temperature estimates. Participants whose mean temperature estimates were two or more standard deviations from the mean estimate were removed from subsequent analysis on the

grounds that they were not taking the task seriously. Additionally, participants who salted on every trial, presumably to guarantee the minimum cash prize in conditions in which that was possible, were removed from analysis. (In the version of the task described in the previous chapter, salting on every trial would not result in cash payment, so in previous analyses, participants who salted on every trial were not removed.) Based on these criteria, a total of 34 participants were removed, leaving 354 participants (56.2% female) for subsequent analysis.

To explore the effect of false alarms and uncertainty estimates on participants' task performance, analyses were performed on measures of decision quality, trust, and compliance. Comparisons were made within each level of forecast format and in comparison to the control. Decision quality was measured by mean expected decision value, which was calculated in the same manner as it was previously. Participants' decisions to salt were evaluated irrespective of the observed temperatures: Applying salt was calculated as having an expected value of -$1,000 and not applying salt was valued at the possible penalty, -$5,000, multiplied by the probability of freezing on that trial. A mean expected decision value was then calculated for each participant. Trust was measured with participants' trust ratings on the 6-point scale (from 1-not at all to 6-completely). The two end-of-month trust ratings were averaged together for a single mean trust rating. Finally, compliance was measured with a compliance score, the compliance ratio, which was calculated by comparing decisions made by participants with decision advice to decisions made by the average control participant (without advice). This made apparent the influence of the advice above and beyond participants' decisions which could have aligned with normative decisions even without the explicit advice of the DSA. A compliance ratio of 1.00 meant that on average, participants were not affected by the advice, i.e., they made the same decisions participants without the advice made; a compliance ratio greater than 1.00 meant that on average,

participants were more compliant, i.e., they made decisions consistent with the advice more than participants without the advice did. Note that a higher compliance score is not *better*, per se, as it simply tracks how closely participants follow the given advice, whether the advice is good or bad.

The effects of false alarm level and forecast format were explored. Over all participants (except controls), as false alarm level decreased, decision quality increased, but at the highest false alarm level, decision quality began to decrease again. As false alarm level decreased, trust increased; and Unadjusted FA-20 and Low FA-30 participants were more compliant than High FA-10 and Low FA-40 participants. Concerning forecast format, participants whose forecasts included both advice and freeze probability estimates demonstrated higher decision quality, greater trust, and higher compliance than participants whose forecasts included only advice. A two-way ANOVA on mean expected decision value, mean trust rating, and compliance score, with false alarm level (High FA-10, Unadjusted FA-20, Low FA-30, and Low FA-40) and forecast format (advice alone and advice + freeze probability) as the independent variables, revealed significant main effects of false alarm level on mean expected decision value, $F(3, 315) = 6.82$, $p < .01$, suggesting that decision quality improved as false alarm level decreased, but only to the 30% threshold level, after which (at the 40% level) it worsened. The ANOVA revealed a significant effect on trust, $F(3, 315) = 3.24$, $p = .02$, with trust increasing as false alarm level decreased. The ANOVA also yielded a strong effect on compliance, $F(3, 315) = 12.52$, $p < .01$, with Unadjusted FA-20 and Low FA-30 participants demonstrating significantly greater compliance than High FA-10 ($p = .01$ and $p < .01$, respectively) and Low FA-40 participants ($p < .01$ and $p < .01$, respectively). There were significant main effects of forecast format on mean expected decision value, $F(1, 315) = 47.10$, $p < .01$, trust, $F(1, 315) = 4.90$, $p =$

.03, and compliance, $F(1, 315) = 7.54$, $p < .01$, suggesting higher decision quality, greater trust, and higher compliance, respectively, among participants who received advice and uncertainty estimates compared to participants who received only advice. See Table 3.4.

Table 3.4. Mean (standard deviation) expected decision value, mean trust, and compliance score by condition.

| | **Mean expected decision value** | **Mean trust rating (6-point scale)** | **Compliance score** |
|---|---|---|---|
| **Advice alone (without uncertainty estimates)** | | | |
| High FA-10 | -$1,084.01 ($83.73) | 2.29 (.87) | 1.05 (.23) |
| Unadjusted FA-20 | -$1,049.55 ($81.97) | 2.42 (1.07) | 1.11 (.17) |
| Low FA-30 | -$1,033.36 ($64.47) | 2.69 (.80) | 1.10 (.15) |
| Low FA-40 | -$1,042.07 ($52.89) | 2.57 (.97) | 0.99 (.26) |
| *mean* | *-$1,052.18 ($73.22)* | *2.49 (.94)* | *1.06 (.21)* |
| **Advice + uncertainty estimates** | | | |
| High FA-10 | -$1,027.79 ($76.68) | 2.38 (.89) | 1.06 (.26) |
| Unadjusted FA-20 | -$991.00 ($45.17) | 2.83 (1.01) | 1.20 (.13) |
| Low FA-30 | -$994.42 ($57.03) | 2.77 (.99) | 1.23 (.16) |
| Low FA-40 | -$997.38 ($42.54)) | 2.94 (.94) | 1.00 (.23) |
| *mean* | *-$1,001.99 ($57.56)* | *2.74 (.97)* | *1.12 (.22)* |
| **Fully crossed: all participants (without control)** | | | |
| High FA-10 | -$1,056.28 ($84.66) | 2.33 (.88) | 1.05 (.24) |
| Unadjusted FA-20 | -$1,019.13 ($71.39) | 2.63 (1.05) | 1.16 (.16) |
| Low FA-30 | -$1,012.39 ($63.27) | 2.74 (.90) | 1.17 (.17) |
| Low FA-40 | -$1,019.73 ($52.74) | 2.76 (.97) | 1.00 (.24) |
| *mean* | *-$1,026.45 ($70.20)* | *2.62 (.96)* | *1.09 (.22)* |
| **Control** | -$1,083.51 ($80.25) | 2.37 (.90) | na |

In the analyses above, the trust ratings were not consistent with the compliance ratings. Compliance and trust increased as false alarm level decreased, but at the lowest false alarm level, compliance decreased despite trust increasing. The relationship between trust and compliance, therefore, was unclear. To explore this further, an additional analysis on the trial-by-trial trust ratings was performed. (The previous analyses had used the end-of-month trust ratings.)

Participants' trial-by-trial trust ratings (from 1-not at all to 6-completely) were averaged across two sets of trials: trials on which participants had complied with the decision advice and trials on which they had not complied. This resulted in two mean trust values for each participant. Analysis revealed that trust was higher on trials on which participants complied with the decision advice than on trials in which they did not comply; trust was higher among participants who were given advice and freeze probability estimates than among participants who were given advice alone; and, importantly, mean trust ratings for trials on which participants complied with the advice followed the predicted false alarm effect pattern: Trust increased as false alarm level decreased. A mixed-model ANOVA on mean trial-by-trial trust ratings, with compliance with advice (complied and did not comply) as the within-participants independent variable and false alarm level (High FA-10, Unadjusted FA-20, Low FA-30, and Low FA-40) and forecast format (advice and advice + freeze probability) as the between-participants independent variables, revealed a significant effect for compliance with advice, $F(1, 308) = 172.65$, $p < .01$, suggesting that participants gave significantly higher trust scores on trials on which they complied than on trials on which they did not comply. There was not a significant main effect for false alarm level, but the significant interaction, $F(3, 308) = 3.34$, $p = .02$, suggested that on trials on which participants complied with the advice, trust increased as false alarm level decreased. Finally, there was a significant effect for forecast format, $F(1, 308) = 6.77$, $p = .01$, suggesting significantly higher trust among participants who received advice with a freeze probability estimate than among participants who received advice alone. See Table 3.5.

Table 3.5. Mean (standard deviation) trial-by-trial trust ratings for trials on which participants complied or did not comply with the decision advice by forecast format and false alarm level.

| **Forecast format** | **False alarm level** | **Trust: Complied** | **Trust: Did not comply** | **Overall** |
|---|---|---|---|---|
| Advice | High FA-10 | 2.70 (.97) | 2.48 (.92) | *2.72 (.89)* |
| | Unadjusted FA-20 | 2.75 (.90) | 2.37 (.70) | |
| | Low FA-30 | 2.85 (.80) | 2.54 (.79) | |
| | Low FA-40 | 3.05 (1.10) | 2.58 (.98) | |
| *Overall (advice)* | | *2.85 (.96)* | *2.50 (.85)* | |
| Advice + freeze probability | High FA-10 | 3.06 (.96) | 2.76 (.80) | *3.02 (.93)* |
| | Unadjusted FA-20 | 3.08 (1.02) | 2.66 (.89) | |
| | Low FA-30 | 3.19 (1.06) | 2.52 (.81) | |
| | Low FA-40 | 3.35 (.97) | 2.76 (.92) | |
| *Overall (advice + freeze probability)* | | *3.18 (1.00)* | *2.67 (.85)* | |
| *Overall* | | *3.02 (.99)* | *2.59 (.86)* | *2.88 (.92)* |

The practical goal of the present experiment was to compare two strategies to increase compliance with weather warnings: 1) adding uncertainty estimates and 2) raising the false alarm level. To determine the effect of inclusion of uncertainty estimates, Unadjusted FA-20 participants with and without uncertainty estimates were compared directly. Advice + uncertainty participants demonstrated better decision quality, higher trust ratings, and greater compliance than did advice participants. An independent-samples t-test on mean expected decision value showed a strong effect of inclusion of uncertainty estimates, $t(55.06) = 3.84$, $p < .01$ (adjusted for unequal variances), with advice + uncertainty participants demonstrating higher decision quality. Another t-test on trust revealed a similar marginally significant effect, $t(75) = 1.71$, $p = .09$, again with advice + uncertainty participants indicating higher trust in the forecasts than participants with advice alone. A final t-test on compliance scores yielded an effect of forecast format, $t(67.90) = 2.61$, $p = .01$ (adjusted for unequal variances). Advice + uncertainty

participants were significantly more compliant than advice-alone participants were. See Table 3.4.

Finally, to determine the effect of raising the false alarm level, Unadjusted FA-20 participants (without uncertainty estimates) and Low FA-30 participants (without uncertainty estimates) were compared directly. They did not differ. An independent-samples t-test on mean expected decision value revealed no significant difference, $t(71) = .94$, $p$ = ns. There was no effect of raising the false alarm level on trust, $t(71) = 1.24$, $p$ = ns, nor was there an effect on compliance, $t(71) = .31$, $p$ = ns. See Table 3.4.

**Discussion**

The results, particularly the final two sets of analyses, strongly suggest that including uncertainty estimates in forecasts with decision advice is a more effective way to increase compliance than raising the false alarm level of the advice. Adding an uncertainty estimate improves decision quality, increases compliance, and slightly increases trust, whereas raising the false alarm level has no significant effect on any of these. Furthermore, including uncertainty estimates in forecasts offers users a more accurate reflection of forecasted weather conditions, unlike adjusting the false alarm level, which at best imposes an externally generated risk tolerance on decision makers and at worst potentially misleads them. Means suggest that in some cases (e.g., decision quality), raising the false alarm rate slightly did indeed lead to better task performance, consistent with the prediction (Roulston & Smith, 2004). But the improvements were not statistically significant. Moreover, the improvements resulting from inclusion of uncertainty estimates were substantially greater. The "uncertainty estimate effect" was much stronger than the cry wolf effect.

The experiment also produced extremely interesting false alarm effects. A classic false alarm effect was observed, but only up to the 30% threshold level. As false alarm rate decreased, decision quality and trust ratings increased, as predicted. However, at the 40% threshold level, the lowest false alarm level tested, decision quality and trust began to decline, perhaps due to the effect of misses: The tradeoff for fewer false alarms is more misses, and at the 40% threshold level, perhaps the effect of those misses began to adversely affect participants, both psychologically and in their ability to effectively use the forecast information. The effect of false alarms was most dramatic on the opposite end of the spectrum: Participants in the highest false alarm rate condition did significantly worse than other participants in almost all analyses. *Increasing* the false alarm rate has not been suggested to promote compliance with weather warnings, admittedly, but the results of the present experiment offer rich evidence of an effect of false alarms on weather-related decision making, an effect which, to my knowledge, had not before been produced in a controlled setting.

One interesting finding was that end-of-month trust ratings followed a classic false alarm effect: As false alarm rate increased, trust decreased. However, that pattern did not hold for compliance or decision quality. In other words, participants expressed greater trust in decision advice that generated fewer false alarms, even though the advice led neither to higher decision quality nor to greater compliance. The trial-by-trial trust measure was more closely related to compliance than the end-of-month measure was, with higher trust on trials on which participants complied with the decision advice. Still, the classic false alarm effect was observed on trials on which participants complied: As false alarm level increased, trial-by-trial trust decreased. The higher trust ratings for lower-false-alarm advice might suggest a degree of wishful thinking on

the part of the participants (Harris et al., 2009), and it stands as evidence of how adjusting the threshold for advising precautionary action can mislead people.

In conclusion, false alarm level appears to have an effect on weather-related decision making. High false alarm rate leads to inferior decision making, and decision quality improves as false alarm rate decreases, but only up to a point, past which decision quality suffers. More importantly, this experiment provides a clear answer to the question of how to increase compliance with weather warnings. The demonstrated benefit of inclusion of uncertainty estimates has critical practical implications. It is a simple (and truthful) adjustment that can be made to weather warnings that might lead to better decision making.

## Chapter 4: Making Decisions About Rare Weather Events

### Background

In the previous set of experiments, it was demonstrated that inclusion of probabilistic uncertainty estimates resulted in an overall improvement in participants' decision quality and an increase in their trust in the forecasts. Furthermore, inclusion of uncertainty estimates was a more effective means of promoting compliance than lowering the false alarm level of decision advice. Overall, this is a compelling line of experimental evidence that non-expert end-users can make effective use of uncertainty estimates and that the public might benefit from their inclusion in weather forecasts. However, the uncertainty estimates were not uniformly helpful. Consider participants' salting decisions in the decision advice and forecast error experiments. Specifically, at low probabilities for which the economically optimal choice was to apply salt treatment, i.e., approximately 17% to 23% probability of freezing, participants who were shown forecasts that included freeze probability estimates actually salted less than participants who were shown only single-value forecasts. They made inferior decisions; in that range of probability, participants were better off with conventional single-value forecasts. This suggests that in situations in which a forecast for an unlikely event still merits taking precautionary action, communicating probabilistic information might be to the detriment of end-users' decision quality.

This is of particular interest because this exact sort of situation occurs in real life. Take, for example, the Halloween Nor'easter of October 2011. On October 28, forecasting models suggested an extremely rare snow storm for much of the U.S. eastern seaboard (Hamrick, 2011). October snow would be an extremely rare event. In New York City, for instance, there had been no measurable October snow since 1952 (Forer, Tanglao, & Schabner, 2011). However, the

forecast was highly uncertain; the likelihood of the storm actually occurring was low. Over the next few days, from October 29 to 31, parts of the mid-Atlantic U.S. to eastern Canada received up to 32 inches of snow, with long-held records broken in at least 20 cities (Hart & Polson, 2011). The issue faced by forecasters in this and other similar situations is how to effectively communicate the risk of rare but extreme weather events to the public. Forecasters want to highlight that although the event has low likelihood of occurring, the potential consequences are serious enough that precautionary action is warranted.

What information could be communicated to people to accurately portray a low-probability event that is extreme enough to merit precautionary action? Looking back to the road salting experiment results, it might have been the inclusion of an uncertainty estimate that led to inferior decisions at relatively low probabilities. On the other hand, it might have been that the *particular type* of uncertainty estimate used, freeze probability, was inappropriate for the task. Taking precautionary action in such a low range of probabilities might have seemed counterintuitive to participants. Indeed, on average, at that level of probability, a freezing temperature was observed on only about one in five trials. However, because of the cost structure of the task, action was warranted at those low probabilities. With a high potential loss, like the potential for a devastating car collision during a snow storm or even loss of life, precautionary action is warranted even when the probability of the event is low. What expression of risk can communicate that and promote precautionary action?

It is possible that communicating an odds ratio, rather than a probability, would effectively communicate the significance of a low-likelihood event. In the domain of weather, an odds ratio expresses the odds of a weather event on a particular day, or forecast odds, relative to the odds of the event over many years of observations, or climatological odds (Murphy, 1991).

Odds ratios express the likelihood of a particular weather event relative to past events, whereas probability estimates simply express the likelihood of a particular weather event occurring ("forecast probability," Zhu & Toth, 2001). In the case of the October Nor'Easter, for example, it might have been the situation that the forecast probability was low – and communicating that to the public would not likely generate urgent response – but the likelihood of the storm at that particular time *relative to* the likelihood of such a storm in a typical October would surely bring greater attention to the matter. In other domains, such as health and medicine, odds ratios and other expressions of relative risk have been used to highlight increases in risk due to a drug or procedure in order to generate concern, rather than expressing the absolute risk, which might be very low (Malenka et al., 1993; Lipkus, 2007). For example, by one estimate, the risk of getting lung cancer from smoking cigarettes is only about 1.25% (Ando et al., 2003), an absolute risk that hardly seems dangerous. Indeed, for activities that are familiar, like smoking, people might discount or ignore a low-probability event, like getting a smoking-related disease (Slovic et al., 2004). Smoking is regarded as relatively controllable and observable (Slovic, 1987). However, if one considers that the risk of getting lung cancer among non-smokers is only about 0.22% (Ando et al., 2003), that means that smokers are almost *six times more likely* to get lung cancer than non-smokers are. Expressing the risk of smoking in relative rather than absolute terms highlights its impact. Furthermore, experimental evidence suggests that expressing relative risk estimates leads to precautionary action (e.g., Stone et al., 1994). A potential pitfall of expressing the risk of rare events as odds ratios is that the expressions are too persuasive: Odds ratios can lead to overcautious decisions, such that people take precautionary action when it is not warranted (Edwards et al., 2001; Likpus, 2007). Thus, odds ratios might not improve decision quality, per se, but rather they might merely lead to more cautious decisions. These effects of odds ratios

might also apply in the context of forecasts for rare but extreme weather events. To my knowledge, no experimental research has yet explored this.

To test the hypothesis of the effectiveness of odds ratios in conveying low-probability but potentially very dangerous weather events, I designed an experiment that used the road salting paradigm. Participants, playing the role of road maintenance manager, again had to make road salting decisions to guard against icy roads resulting from freezing temperatures, but in addition they had to decide if they wanted to guard against a rare, extreme event: a temperature plunge to 0°F or below, which would necessitate a special sort of salt treatment. I systematically manipulated the likelihood of the rare event between conditions, so, based on the cost-loss structure, for some participants precautionary action was warranted, whereas for others it was not. I also manipulated the way the forecast information was communicated. Some participants were shown only single-value forecasts. Some participants' single-value forecasts were augmented with freeze probability estimates, and others' were augmented with odds ratios expressing the likelihood of the extreme event on that night relative to such an extreme event in the climatological record. I hypothesized that participants shown odds ratios would demonstrate the greatest precautionary action for the extreme cold-weather event, but that participants shown freeze probability estimates would do best overall, consistent with the previous experiments.

For clarification, it should be noted that in this experiment, rareness and extremeness are distinct concepts. A rare event is not necessarily extreme, although an extreme event is, by definition, rare (Palmer & Räisänen, 2002). More importantly, in the context of the present experiment, the rareness and extremeness of the extreme low temperature target event were treated separately. The rare event of interest in this experiment was a temperature of 0°F or colder, an event which, according to climatological data from Washington State, is very rare. The

likelihood (or forecast probability) of that rare event was systematically manipulated across conditions in this experiment. Extremeness, on the other hand, was operationalized as the cost-loss ratio, specifically the denominator of that ratio, the loss. Extremeness was held constant across conditions (0.17); it was not manipulated in the present experiment. To manipulate extremeness, one would systematically vary the potential loss between conditions, e.g., a special salt treatment cost of $2,000 and potential penalties of $10,000, $20,000, and $40,000, characterizing ever more extreme events with cost-loss ratios of 0.20, 0.10, 0.05, respectively. This maps precisely onto the real-life situation: A more extreme event presumably would incur a greater potential loss. Therefore, in referring to "rare, extreme events" in the present experiment, "rare" specified that the event was indeed climatologically rare and was forecasted to occur at varying levels of probability. "Extreme," on the other hand, was simply additional information that highlighted what this experiment was designed to explore: rare events that are extreme, like October snow storms. The extremeness was the same across all levels of the rareness variable. Future research might explore the relationship between rareness and extremeness (perhaps with a fully crossed design in which both are manipulated) and what forecast format most effectively elicits economically optimal decision making.

**Method**

For the rare event experiment, the basic design of the road salting task was modified to include an additional response option. In this version of the task, participants chose either to withhold salt or apply salt treatment to guard against freezing temperatures, as before, or they could apply a special salt treatment that guarded against extreme low temperatures of 0°F and below. As before, regular salt treatment for freezing temperatures cost $1,000 per application, and failing to apply treatment when a freezing temperature was observed resulted in a penalty of

$6,000. Special salt treatment cost $2,000, and failing to apply treatment when a sub-0°F temperature was observed resulted in a penalty of $12,000. (Using special salt when a temperature between 1°F and 32°F was observed protected against any penalty but was inefficient in that it cost more than an application of regular salt, which was all that was necessary to guard against a freezing temperature. Using regular salt when a sub-0°F temperature was observed resulted in the $12,000 penalty, plus the cost of the regular salt treatment.) Therefore, the cost-loss ratio for both types of decisions, i.e., using regular salt to guard against freezing temperatures and using special salt to guard against sub-0°F temperatures, was one-sixth, meaning that participants ought to have applied regular treatment when the probability of a freezing temperature was 17% or greater, and they ought to have applied special salt when the probability of a sub-0°F temperature was 17% or greater.

Participants. Two hundred ninety-four participants (48% female) participated in the experiment. Age ranged from 18 to 32 years ($M$ = 19 years). All participants were recruited from introductory psychology courses at the University of Washington and were awarded a small amount of academic credit for participating, plus a small cash incentive commensurate with their performance. Participants signed up voluntarily.

Apparatus. The experiment was programmed with Microsoft Visual Basic and was administered on standard desktop computers.

Procedure. Participants arrived in groups of 1 to 12 at pre-organized times in a computer lab. After participants read a consent form and agreed to participate, the experimenter presented task instructions and a practice trial while participants followed along on their computers. Following an opportunity for questions, the task began. Participants completed the task individually. Upon

completion of the task, participants answered some follow-up questions and were dismissed. All participants who showed up for the experiment were awarded credit.

Stimuli. A new forecast set of 60 trials was created for this experiment that featured a severe drop in temperature over a period of several days late in the first month. This temperature drop led to the extreme low temperature target trial. See Table 4.1.

Table 4.1. Forecast characteristics.

| | **Forecasts in original road salting experiment (low-error trials)** | **Forecasts in rare event experiment (except target and trial before)** |
|---|---|---|
| **Mean forecast error** | 3.26°F | 3.17°F |
| **Mean freeze probability** | 24.75% (range 10-51%) | 35.31% (range 10-100%) |
| **Mean forecast** | 34.00°F (range 32-37°F) | 32.66°F (range 12-37°F) |
| **Mean observation** | 35.00°F (range 27-42°F) | 32.90°F (range 10-41°F) |

A critical goal of the rare event experiment was to test the effectiveness of weather uncertainty communicated as relative risk, specifically as odds ratios. An odds ratio is essentially a ratio of ratios, comparing the odds of one event to the odds of another. Here I was interested in the ratio of forecast odds, i.e., the odds of a temperature occurring on a particular night, to climatological odds, i.e., the historical odds of a temperature occurring during a period of time. The experiment included forecasts for both freezing and extreme cold (sub-0°F) temperatures. Therefore, two odds ratios were calculated for each trial of the experiment: the odds of a freezing temperature (1°F to 32°F) being observed on that particular trial and the odds of an extreme cold temperature (0°F or colder) being observed on that particular trial (Murphy, 1991). Approximately 100 years of climatological data from the University of Washington's Department of Atmospheric Sciences was used to calculate the climatological odds, specifically

using data from Yakima and Spokane, WA, from November through February. Forecast odds for each trial were the ratio of the likelihood of a temperature below a threshold to the likelihood of a temperature above the threshold. For example, for a 35% probability of a temperature at or below 0°F, forecast odds = .35/.65 = .538. Climatological odds were the ratio of historical occurrences of a temperature below a threshold to occurrences of a temperature above the threshold. For example, if there were 400 occurrences of temperatures at or below 0°F out of 10,000 total observations, climatological odds = 400/9,600 = .042. The odds ratio, therefore, was the ratio of forecast odds to climatological odds, in this example .538/.042 = 12.81. In the experiment, that value would be expressed as follows: "Compared to a typical winter night, the odds are 12.81 times greater tonight that the temperature will be ≤ 0°F." For each of the four levels of the likelihood of rare event variable, odds ratios were calculated to match the probabilities of 0°F-or-below temperature on the extreme low temperature target trial. See Table 4.2.

Table 4.2. Single-value forecasts, probabilities, and odds ratios for the target trial at each level of the likelihood of rare event (0°F) variable.

| **Likelihood of rare event (0°F) level** | **Single-value forecast** | **Probability of ≤ 0°F** | **Odds of temperature ≤ 0°F greater than typical winter night** |
|---|---|---|---|
| 10% level | 6°F | 10% | 3.5x |
| 17% level | 5°F | 17% | 6x |
| 24% level | 3°F | 24% | 9.5x |
| 31% level | 2°F | 31% | 13.5x |

Design. The experiment used a fully crossed 3 x 4 between-participants design. Participants were randomly assigned to one of 12 conditions. The two independent variables were forecast format and likelihood of rare event. The forecast format manipulation had three levels: probability, odds ratio, and control. See Table 4.3.

Table 4.3. Example forecasts for each forecast format.

| **Forecast format** | **Example forecast** |
|---|---|
| Probability | "The expected nighttime low temperature is 35°F; there is a 20% chance the temperature will be 32°F or less and a 0% chance it will be 0°F or less." |
| Odds ratio | "The expected nighttime low temperature is 35°F. Compared to a typical winter night, the odds are 1.5 times greater tonight that the temperature will be 32°F or less and no greater that it will be 0°F or less." |
| Control (single value) | "The expected nighttime low temperature is 35°F." |

The likelihood of rare of event manipulation had four levels: probability of extreme low temperature (0°F) of 10%, 17%, 24%, and 31%. For this manipulation, a single trial, called the extreme low temperature target trial, was different in each condition. At the 10% level, the probability of a sub-0°F temperature was 10%; at 17%, the probability was 17%; and so on. Note that based on the cost-loss ratio, participants ought to have applied special salt treatment on the extreme low temperature target trial at the 17%, 24%, and 31% levels, but not at the 10% level. The trial that immediately preceded the target trial, which was the same across conditions, had a relatively high probability of 0°F or below (5%), but special salt was not warranted on that trial. Therefore, the forecast sequence was identical across all four levels *except* for the extreme low temperature target trial.

**Results**

It was predicted that participants whose forecasts included probability estimates (for both 32°F and 0°F) would do best at the task overall (non-target trials), consistent with the results of the previous experiments. For the rare weather event (target trial), however, participants with odds ratios were predicted to make better decisions, although they were expected to overestimate risk when the risk of the rare event was very low. First, like with the experiments reported in the

previous chapter, participants whose data suggested they did not understand the task or were not taking it seriously were removed by comparing their temperature estimates to the forecasts, and participants with mean error estimates that were two or more standard deviations above the mean standard error for temperature estimates in their experimental condition were removed. Eight such participants were dropped, leaving 286 participants for subsequent analysis.

To explore the hypotheses, a series of analyses were performed separately on the non-target trials and on the target trial alone. Non-target-trials analyses were conducted to evaluate decision quality overall, and target-trial analyses were conducted to evaluate decision quality specifically for the extreme cold temperature target trial.

Participants who were given forecasts with probability estimates demonstrated higher decision quality overall than participants with odds ratios or single-value forecasts alone. Mean expected decision values were calculated for all participants in the manner described in the decision advice and forecast error experiments. For non-target trials, the decision of interest was whether participants applied regular salt treatment (cost of $1,000) on each trial or risked a penalty (potential loss of $6,000). As such, for each trial and for each participant, expected decision values were entered as -$1,000 for decisions to salt and -$6,000 multiplied by the freeze probability on that trial for decisions to withhold salt. Expected values were averaged per participant. A one-way ANOVA on mean expected decision value, with forecast format (probability, odds ratio, and control) as the independent variable, revealed a significant effect, $F(2, 285) = 7.33$, $p < .01$. Tukey's post hoc analysis showed that probability participants made significantly better decisions than did odds ratio participants, $p < .01$, and control participants, $p = .05$. See Table 4.4.

Probability participants also trusted the forecasts more than other participants did. The two end-of-month trust ratings were averaged and analyzed, and an ANOVA on mean trust rating was significant, $F(2, 285) = 4.70$, $p = .01$. Tukey's post hoc analysis revealed that probability participants gave significantly higher trust ratings than odds ratio participants did, $p = .01$, and marginally significantly higher ratings than control participants did, $p = .06$. See Table 4.4.

Finally, I analyzed participants' ability to discriminate between situations in which applying salt was and was not warranted by analyzing the proportion of times they applied salt above the 17% cost-loss threshold (when it was warranted) and below the 17% threshold (when it was not warranted). Overall, participants successfully made that discrimination, but probability participants did it best. A mixed-model ANOVA, with probability of freezing (greater than 17% and less than 17%) as the within-participants variable and forecast format as the between-participants variable, yielded a significant effect for freeze probability, $F(1, 283) = 4{,}328.59$, $p < .01$, with participants applying treatment more often above the 17% threshold than below it. There was an effect for forecast format, as well, $F(2, 283) = 3.29$, $p = .04$, with Tukey's post hoc analysis suggesting that the effect was driven mostly by probability participants salting less overall compared to control participants, $p = .04$. Finally, there was a significant interaction, $F(2, 283) = 12.55$, $p < .01$, suggesting that compared to odds ratio ($p < .01$) and control ($p = .05$) participants, probability participants were better able to discriminate between conditions that warranted treatment and those that did not. See Table 4.4.

Table 4.4. Mean (standard deviation) expected decision value, mean trust rating, and proportion of trials on which participants salted on non-target trials.

| **Forecast format** | **Mean expected decision value** | **Mean month trust rating** | **Salt decisions (proportion)** | |
|---|---|---|---|---|
| | | | **< 17%** | **≥ 17%** |
| Probability | -$1,155.67 ($118.63) | 2.58 (.81) | .08 (.15) | .66 (.13) |
| Odds ratio | -$1,236.84 ($157.26) | 2.27 (.71) | .14 (.17) | .63 (.15) |
| Control (single value) | -$1,206.61 ($161.48) | 2.33 (.69) | .17 (.18) | .67 (.16) |
| *mean* | *-$1,200.30 ($150.61)* | *2.39 (.75)* | *.13 (.17)* | *.65 (.15)* |

Lastly, decisions on the extreme cold temperature target trial were analyzed, revealing that odds ratio participants took precautionary action (applied special salt) more than other participants at all levels of target event likelihood. To explore whether participants took appropriate precautionary action on the target trial, I conducted a binary logistic regression on the application of special salt, coded as a binary variable, with two independent variables: forecast format (probability, odds ratio, and control) and probability of 0°F (10%, 17%, 24%, and 31%). Results suggested that odds ratio participants were 2.29 times more likely than control participants to use special salt, Exp(*B*) = .44, $p = .01$, and 14.49 times more likely than probability participants, Exp(*B*) = .07, $p < .01$. Results also revealed an effect of probability of 0°F, showing that participants were more likely to apply special salt above the 17% threshold than below it. Compared to participants for whom the probability of 0°F was 10% on the target trial, participants in the 17% condition were 1.93 times more likely to apply special salt, Exp(*B*) = 1.93, $p = .09$ (marginally significant effect); participants in the 24% condition were 5.95 times more likely to apply special salt, Exp(*B*) = 5.95, $p < .01$; and participants in the 31% condition were 8.72 times more likely to apply special salt, Exp(*B*) = 8.72, $p < .01$. Finally, there was a significant interaction between forecast format and probability of 0°F, Wald's $\chi^2(6) = 34.39$, $p < 0.01$, indicating that the probability of 0°F differently affected forecast format conditions. Odds

ratio participants, for instance, applied special salt more at high probabilities of 0°F than at low probabilities, whereas probability participants applied special salt at approximately equal levels across probabilities of 0°F. See Table 4.5.

Table 4.5. Proportion of participants (standard deviation) who applied special salt on the target trial.

| **Forecast format** | **10%** | **17%** | **24%** | **31%** | *mean* |
|---|---|---|---|---|---|
| Probability | .15 (.36) | .17 (.39) | .18 (.39) | .29 (.46) | .19 (.40) |
| Odds ratio | .41 (.50) | .65 (.49) | .91 (.29) | .91 (.29) | .70 (.46) |
| Control (single value) | .28 (.46) | .38 (.49) | .74 (.45) | .83 (.38) | .55 (.50) |
| *mean* | *.24 (.43)* | *.37 (.49)* | *.59 (.49)* | *.75 (.43)* | *.49 (.50)* |

**Discussion**

The results clearly suggest that the perceived risk of a rare event is greater when the risk is communicated as an odds ratio than as a probability. When threatened with non-rare events (i.e., freezing temperatures in winter months), participants shown odds ratio forecasts took no more precautionary action than did participants shown only single-value forecasts, but they took significantly more precautionary action than other participants when threatened with a single rare event (i.e., temperature of 0°F or below). Participants given probability estimates, on the other hand, made the most economically optimal decisions overall when threatened with non-rare events, but they did not exercise sufficient precautionary action when threatened with a single rare event.

Together, these results strongly suggest that certain forecast expressions are more suitable for some situations than for others. Odds ratios seem a particularly appropriate means to

communicate rare, extreme events, but they are inappropriate for communicating non-rare events or typical weather. Indeed, for the non-target trials, the additional odds ratio figure provided no additional information to odds ratio participants, as the odds of freezing temperatures were usually "no greater than on a typical winter night." The information was useless to participants. Conversely, probability estimates are best used for typical weather events but not for rare events. Interestingly, the probability participants actually took less precautionary action on average for the extreme low temperature event than did participants who had no information beyond just the low temperature forecast. This underscores that in some cases, inclusion of probabilistic uncertainty estimates not only does not benefit decision makers, but it can actually hurt them.

Forecasts with odds ratios can potentially mislead decision makers in some situations, as well. While the experimental participants who were shown odds ratios demonstrated significantly more appropriate precautionary action than other participants did, they also demonstrated significantly more *inappropriate* precautionary action. More than two-fifths of odds ratio participants applied special salt when the likelihood of the extreme cold temperature was not sufficiently high to warrant doing so. The odds ratio format seemed to amplify the risk of the event, making all risks – trivial and significant alike – seem important. Indeed, that is arguably the point of the odds ratio format, but those using it to communicate risk should only do so when they are attempting to persuade a particular course of action (Lipkus, 2007). The results of the experiment confirm the concern that odds ratios lead to overestimation of risk (Edwards et al., 2001). The overarching conclusion, then, is that there is not a one-size-fits-all risk communication format for weather-related risk. Forecast providers need to take great care in choosing the forecast expression that is most suitable for the given situation.

# Chapter 5: Communicating Climate Change Uncertainty

## Background

The previous experiments provide evidence that the inclusion of uncertainty estimates in weather forecasts can lead to better decision making. However, it remains unclear if the benefit conferred to decision makers in weather-related decision making extends to other types of decisions. One area of pressing interest is how people understand climate change, including how they perceive climate-related uncertainty and how they make climate-related decisions. In this chapter, I will explore perception of climate-related uncertainty and ways to communicate climate change uncertainty that might enhance people's belief that the climate is changing.

More Americans than ever believe the global climate is changing, but a substantial minority of skeptics still exists. Despite the near complete consensus among climate scientists that climate change is occurring and that human activity is the cause (Cook et al., 2013), a recent survey in the United States found that fully 36% of those polled either did not believe in climate change or were unsure (Leiserowitz et al., 2013). Perhaps partly because of the conviction of that minority, significant large-scale action to mitigate climate change has still failed to occur. Broader belief in climate change, or a closer match in the degree of belief between the scientific community and the public, is necessary to drive substantive action to combat climate change.

Why people still do not believe that climate change is occurring is the subject of sprawling debate. Belief (or lack of belief) in climate change is an extraordinarily complex topic, and numerous interconnected factors likely determine where one stands (see CRED, 2009, for a review). One possible factor is that people believe climate scientists are conflicted about climate change (Ding et al., 2011). A recent survey revealed that 33% of respondents in the United States

believed there is "a lot" of disagreement among scientists about whether global warming is occurring or not (Leiserowitz et al., 2013). Skeptics might misinterpret variability among scientists' estimates and debate within the climate science community as an indication that climate scientists are dishonest or incompetent (Johnson & Slovic, 1995; Johnson & Slovic, 1998) or that there is not scientific consensus that climate change is occurring at all (Freudenburg & Muselli, 2010). The perceived disagreement could then lead to distrust in the projections (Moser, 2010). This misperception is further bolstered by the news media, which, in an attempt to present a balanced story, tend to characterize debate within the science community as fundamental disagreement and overemphasize reports of the tiny minority of opposing scientists (Freudenburg & Muselli, 2010; Boykoff, 2009).

Improving the communication of climate change projections could make them more convincing. One possible communication improvement would be to add an estimate of climate uncertainty, which would perhaps allow skeptics to make sense of the disagreement among climate scientists and explain the variation in their projections. While the idea of communicating probability or a range of values might at first seem like it would undercut the message, perhaps by making the scientists seem unsure or incompetent, mounting evidence on how non-expert laypeople perceive risk suggests otherwise. As demonstrated in the road salting experiments reported earlier and in other research (Nadav-Greenberg & Joslyn, 2009; Roulston et al, 2006; Morss et al., 2008), the inclusion of uncertainty estimates in weather forecasts is beneficial to end-users, in that it increases users' trust and leads to more economically optimal decision making.

The effect of increased trust in weather forecasts that include uncertainty estimates is plausibly due to two factors. One factor is that observed temperatures are much more likely to

fall within a range of possible temperatures (which is expressed by the uncertainty estimate) than to exactly match a single-value forecast. That is, forecasts with uncertainty estimates likely seem less inaccurate if they fail to verify. The other factor is that people have largely correct intuitions about weather uncertainty. Research has revealed that people correctly intuit that there is less certainty as lead time increases (Morss et al., 2008) and that extreme events are rare (Joslyn & Savelli, 2010). More generally, many people infer ranges of values around single-value estimates even when no ranges are stated explicitly (Johnson & Slovic, 1995). Therefore, forecasts that include uncertainty estimates might seem more realistic, complete, and trustworthy.

The benefit to users of having uncertainty estimates might not be limited to weather forecasts. It is possible that including ranges of values in climate projections will lead people to trust them more than single-value projections. However, climate projections are in many ways different from weather forecasts. Most notably, climate projections concern far-future events, events perhaps beyond the users' lifetimes, such that people cannot develop an understanding of the relationship between projections and outcomes. In contrast, with weather forecasts, the forecast-observation relationship has been developed throughout the users' lifetimes. As such, the advantage of uncertainty expressions to users of weather forecasts – who can compare the probabilistic forecasts with observed weather – might not extend to climate projections. However, if people have good intuition about the uncertainty inherent in climate projections, then projections that include ranges of possible values might seem more realistic and therefore more trustworthy than single-value projections. The present research explores the inclusion of ranges of possible values in climate projections to see if this is indeed the case.

I created a simple experiment in the form of a questionnaire to administer online. The questionnaire presented research participants with a pair of climate change projections and asked

them to give their own estimates of the future parameter values, plus ratings of trust and concern in the projections. Some participants were shown only conventional single-value climate projections (e.g., a projected temperature increase of 3°F); other participants were shown projections that included a range of values around a most likely single value. Additionally, I manipulated the target year of the projections to explore people's perception of climate uncertainty over time. I was not sure what to expect, as there is compelling reason for opposite hypotheses concerning the effect of inclusion of uncertainty estimates in the climate projections. If the ability to compare forecasts with outcomes is an essential aspect to the advantage of including uncertainty estimates in weather forecasts, then it could be expected that including uncertainty estimates in climate projections would not have any effect. However, if participants have good intuition about the uncertainty inherent in climate projections, then it could be expected that participants who were shown uncertainty estimates would judge the projections more trustworthy, resulting in higher ratings of concern and personal estimates of future climate parameters that were closer to the projected values they were shown.

Data were obtained from two populations: a pilot sample of visitors of a Pacific Northwest-based weather blog, and a primary sample of users of Amazon's Mechanical Turk, an online research tool. The questionnaire was slightly modified between samplings. The specific design and stimuli for each will be described in turn below.

## Experiment 1: Weather Blog Sample

### Method

Participants. The questionnaire was posted in electronic format on a popular Pacific Northwest U.S. weather blog. Over a period of approximately two weeks in August 2012, 818 people (30.2% female) completed the questionnaire.

Procedure. A link to the questionnaire was posted on an internet blog about weather and climate. Participation was optional. Prospective participants were shown an information sheet about the research and their rights as participants before they chose to participate or not. There was no compensation for participation.

Stimulus and Measures. The questionnaire presented respondents with two climate projections: an average temperature increase of 3°F and an average sea level rise of 6 inches over 20$^{th}$ century averages by 2050. These values were based on climate reports and on values obtained using the MAGICC/SCENGEN program, a free, downloadable computer program that allows users to create custom projections based on different climate models. After each projection, respondents answered five questions: their own parameter estimate (e.g., "I think the increase in average global temperature by 2050 will be ___°F"), a lower-bound estimate (e.g., "I would not be surprised if the increase in average global temperature by 2050 were as little as ___°F"), an upper-bound estimate (e.g., "I would not be surprised if the increase in average global temperature by 2050 were as great as ___°F"), a rating of trust in scientists' estimate (on a 6-point scale, from "not at all" to "completely"), and a rating of concern about the projection (on a 6-point scale, from "not at all" to "extremely").

Design. The questionnaire had a 2 x 4 between-participants factorial design. One independent variable, projection format, had two levels: deterministic and uncertainty. In the deterministic format, single-value projections were shown for temperature and sea level; in the uncertainty condition, the projection values were accompanied by a range of possible values, which represented the upper and lower bounds of a 90% confidence interval. In this uncertainty condition, the temperature projections were for an increase of 3°F over the 20$^{th}$ century average temperature, with a 90% chance the increase would be between 2°F and 4°F; the sea level projection was for a rise of 6 inches over the 20$^{th}$ century average sea level, with a 90% chance the rise would be between 3 inches and 9 inches. The uncertainty estimates were also obtained using the MAGICC/SCENGEN program. Although confidence intervals are not yet reliably calculatable for climate projections, I used theoretical interval values here because such intervals are a relatively common way to express uncertainty in a number of domains and because it is a goal of climate scientists to provide these estimates. Participants were randomly assigned to one of the two projection formats. The other independent variable was time, which had four levels: 2015, 2025, 2050, and 2100. Participants were told that the climate projections (temperature increase of 3°F and an average sea level rise of 6 inches) were for the year 2015, 2025, 2050, or 2100. Parameter (temperature and sea level) could be considered a within-participants independent variable, but I did not have any hypotheses concerning the relationship between responses to the temperature and sea level items.

The questionnaire was programmed using a simple online questionnaire tool provided by the University of Washington, which was then posted to a weather blog. The Internet link posted on the weather blog was programmed to alternate between the two conditions in order to ensure an approximately equal number of respondents in each condition.

## Results

It was hypothesized that if climate projections were perceived similarly to weather forecasts, inclusion of uncertainty estimates would increase participants' trust in and concern about the projections, and that it would lead to higher and more accurate parameter estimates and estimates of parameter uncertainty. Additionally, it was predicted that participants across conditions would estimate more uncertainty (difference between upper- and lower-bound parameter estimates) as projection year was farther in the future. Before analyses, the data of some respondents were removed. All of a respondent's data were excluded from subsequent analysis if the parameter for either temperature or sea level was two or more standard deviations above the mean parameter estimate or if the lower-bound estimate was greater than the upper-bound estimate (or both). Data cleaning resulted in the removal of 34 participants' data, leaving 784 (30.9% female) for subsequent analysis.

Projection format had a clear effect on trust and concern. Participants who were shown climate projections with uncertainty estimates generally gave higher ratings of trust and concern than participants who were shown only single-value projections. T-tests on trust ratings, with projection format (deterministic and uncertainty) as the independent variable, revealed that uncertainty participants gave significantly higher ratings, both for temperature, $t(782) = 2.12$, $p = .03$, and for sea level, $t(782) = 2.12$, $p < .01$. T-tests on concern ratings revealed that uncertainty participants gave marginally significantly higher ratings only for temperature, $t(782) = 1.87$, $p = .06$. See Table 5.1.

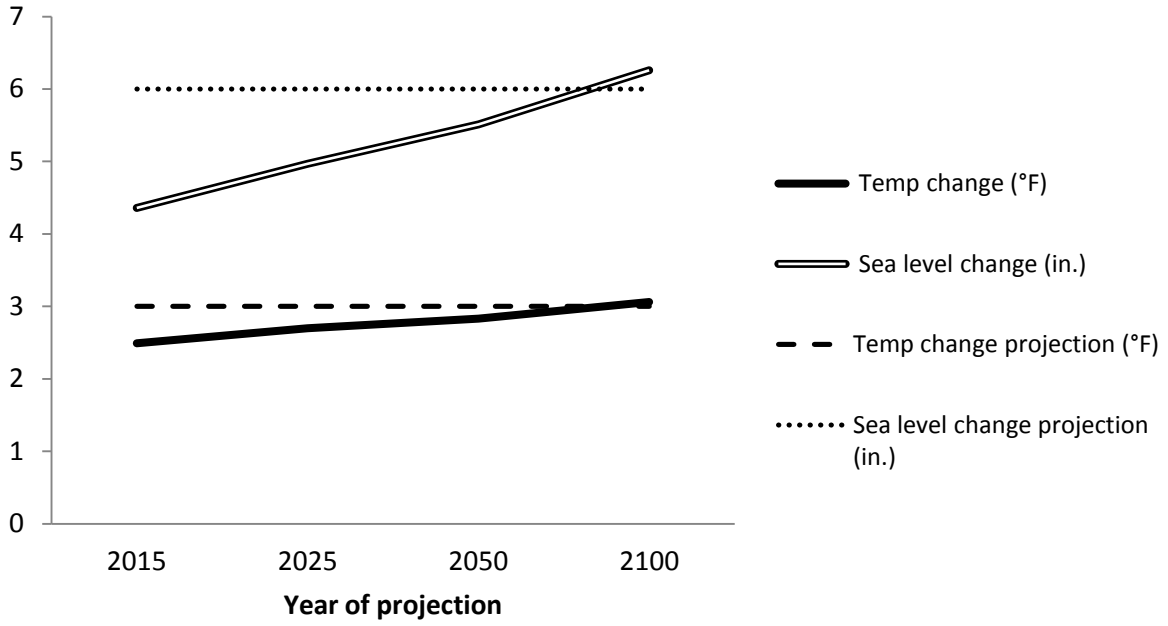
Table 5.1. Mean ratings (standard deviation) by projection format.

| **Projection format** | **Temperature trust** | **Sea level trust** | **Temperature concern** | **Sea level concern** |
|---|---|---|---|---|
| **Deterministic** | 3.62 (1.29) | 3.52 (1.33) | 3.50 (1.58) | 3.57 (1.57) |
| **Uncertainty** | 3.82 (1.32) | 3.77 (1.32) | 3.71 (1.50) | 3.70 (1.52) |
| ***overall*** | *3.72 (1.31)* | *3.64 (1.33)* | *3.60 (1.54)* | *3.63 (1.54)* |

Comparing participants' ratings to the scale midpoint also revealed the effect of inclusion of uncertainty estimates. Participants with uncertainty estimates gave ratings that were significantly above the midpoint, whereas participants with only single-value projections did not. A series of one-sample t-tests on trust and concern ratings, with a criterion value of 3.5, suggested that uncertainty participants gave significantly above-midpoint ratings of trust for temperature projections, $t(383) = 4.77$, $p < .01$, trust for sea level projections, $t(383) = 4.03$, $p < .01$, concern about temperature projections, $t(383) = 2.72$, $p < .01$, and concern about sea level projections, $t(383) = 2.55$, $p = .01$, whereas deterministic participants gave marginally significantly above-midpoint ratings of trust only for temperature projections, $t(399) = 1.90$, $p = .06$

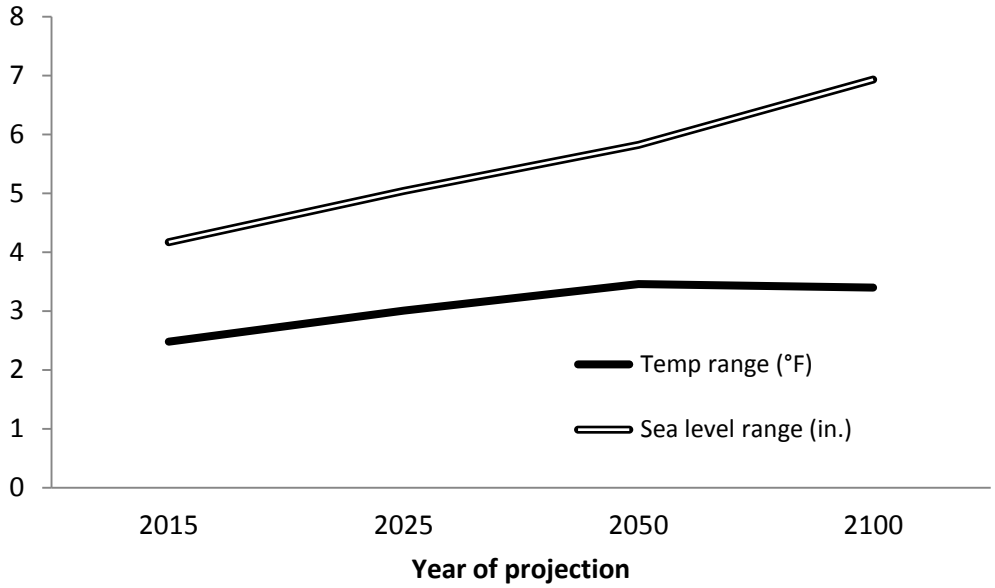
Next, participants' parameter estimates were analyzed. Of 784 participants, 727 (93%) estimated a temperature increase, 10 (1%) estimated a temperature decrease, and 47 (6%) estimated no change; for sea level, 739 (94%) estimated an increase, 1 (< 1%) estimated a decrease, and 44 (6%) estimated no change. On average, participants provided parameter estimates that were lower in value than those in the projections they were shown, even specifically among participants who believed temperature and sea level would rise. See Figure 5.1.

Figure 5.1. Participants' estimates of parameter change among participants who believed temperature and sea level would rise.




Additionally, as the lead time increased (i.e., as the target year for the climate projections went farther into the future), participants' estimates of uncertainty (their parameter ranges, the differences between and lower- and upper-bound estimates) got larger, both for temperature, Pearson's $r = .11$, $p < .01$, and for sea level, Pearson's $r = .24$, $p < .01$. The significant positive correlations suggest that as lead time increased, estimates of uncertainty also increased. See Figure 5.2. The effect of perception of increased uncertainty with time was no different between deterministic and uncertainty participants. T-tests on participants' interval widths, with projection format as the independent variable, were insignificant, suggesting that participants' perception of climate uncertainty was the same regardless of whether they were shown a projection that included uncertainty estimates.

Figure 5.2. Participants' estimates of uncertainty (upper-bound estimates – lower-bound estimates) over time.




Finally, participants were categorized as climate change "believers" or "doubters." Participants who gave temperature change estimates greater than 0°F (meaning any increase in average temperature) were categorized as believers, and participants who gave temperature change estimates of 0°F (meaning no change) or less (meaning decrease in average temperature) were categorized as doubters. With this rough method of categorization, it was revealed that the sample population overwhelmingly believed in climate change (93%). There were slightly more believers in the uncertainty condition than in the deterministic condition, but the difference was not significant, Pearson's chi-square(1) = 1.83, $p$ = .18. See Table 5.2.

Table 5.2. Counts (proportion) of believers and doubters by projection format.

| | **Believers** | **Doubters** | |
|---|---|---|---|
| **Deterministic** | 366 (91.5%) | 34 (8.5%) | 400 (100%) |
| **Uncertainty** | 361 (94.0%) | 23 (6.0%) | 384 (100%) |
| ***Overall*** | *727 (92.7%)* | *57 (7.3%)* | *784 (100%)* |

Not surprisingly, doubters on average gave lower ratings of concern than believers did, both for temperature, $t(77.58) = 16.26$, $p < .01$, and for sea level, $t(69.93) = 12.43$, $p < .01$ (both adjusted for unequal variances), suggesting that the method of categorization described above was legitimate. However, among doubters, participants who were shown uncertainty estimates gave marginally significantly higher temperature concern ratings than deterministic participants did, $t(31.03) = 1.89$, $p = .07$ (adjusted for unequal variances). Sea level concern ratings were higher, as well, but not to statistical significance. See Tables 5.3 and 5.4.

Table 5.3. Mean (standard deviation) concern ratings of believers and doubters.

| | **Believers** | **Doubters** |
|---|---|---|
| **Temperature concern** | 3.77 (1.46) | 1.51 (.97) |
| **Sea level concern** | 3.78 (1.47) | 1.70 (1.19) |

Table 5.4. Among doubters, mean (standard deviation) concern ratings by projection format.

| | **Uncertainty** | **Deterministic** |
|---|---|---|
| **Temperature concern** | 1.83 (1.23) | 1.29 (.68) |
| **Sea level concern** | 1.91 (1.38) | 1.56 (1.05) |

**Discussion**

The projection format hypothesis was supported. Participants who were shown climate projections that included a range of possible values expressed greater trust and concern than participants shown only single-value projections did. Uncertainty participants also gave ratings that were generally high (i.e., above the scale midpoint), whereas deterministic participants did not. Furthermore, even among participants that were characterized as doubters, those who were shown uncertainty estimates gave slightly higher ratings of trust and concern. While the inclusion of uncertainty estimates did not itself *lead to* participants being believers (as evidenced

by the equal proportion of believers and doubters between the projection format conditions), clearly it had some influence on their perception of climate change risk.

Although participants generally underestimated future temperature and sea level compared to the projected values, they inferred higher temperature and sea level farther in the future. Furthermore, the exploration of respondents' perception of climate change uncertainty over time revealed that the farther in the future the climate projection target year was, the greater the uncertainty participants expected, suggesting that they inferred there is greater uncertainty in farther-future events. This is particularly interesting because target year was manipulated between participants. Each respondent gave lower- and upper-bound estimates only for a single year, and he or she had no direct knowledge that other respondents were completing questionnaire versions with different target years. Taken together, the results suggest that people already know there is uncertainty in predicting the future climate. Including an estimate of that uncertainty in the projection might then make it more trustworthy and lead to greater concern. The effect on trust and concern was small, but it is noteworthy in light of the fact that inclusion of uncertainty estimates did not *decrease* trust and concern, as scientists themselves fear (Frewer et al., 2003).

The participant sample used in Experiment 1 was not highly representative of the overall U.S. population. It was disproportionately male (65.6%), it was older than average (mean age: 51.48 years), most respondents were from the same region of the country (Puget Sound area, Washington State: 92.7%), and, very significantly, all respondents were readers of a blog about weather and climate, suggesting that they likely had a richer understanding of the questionnaire items than did the general public and, moreover, a greater inclination to believe in climate change and trust scientists' estimates. Indeed, only 7% of the weather blog sample were

categorized as "doubters." Importantly, Study 1 respondents demonstrated intuitive understanding that climate change uncertainty increases in the future, suggesting that a key piece of the psychological mechanism that might be benefitted by inclusion of uncertainty estimates was indeed there, although it was unknown if that intuition was shared among a broader sample of the public. Therefore, I administered the study a second time, soliciting participation from Amazon's Mechanical Turk, with the aim of obtaining a more representative sample.

## Experiment 2: Mechanical Turk Sample

### Method

Experiment 2 was nearly identical to Experiment 1. However, instead of posting the link to the online questionnaire on a weather blog, which was presumably visited by people with keen interest in weather and climate, I posted the link on Amazon's Mechanical Turk. The M-Turk is a portal to thousands of surveys covering a wide range of topics. Respondents complete surveys for small monetary rewards. Evidence suggests that M-Turk samples are reliable, potentially more internally valid than other data collection sources, and reasonably representative of the general public (Paolacci, Chandler, & Ipeirotis, 2010). Experiment 2 focused solely on the primary goal of Experiment 1, testing the effect of including uncertainty estimates in climate projections for a single target year, 2050. Additional demographic information was collected to ensure that the sample was indeed more representative of the general population.

Participants. The questionnaire was posted in electronic format on Amazon's Mechanical Turk website. Over a period of approximately three months in summer and autumn 2013, 1015 people (56.7% female) completed the questionnaire. Participants' mean age was 33.42 years. In the revised version of the questionnaire, more elaborate demographic information was solicited,

including educational attainment and political party affiliation. See Table 5.5 for comparison of the sample demographic information to that of the general public.

Table 5.5. Comparison of demographics of sample population and general American population.

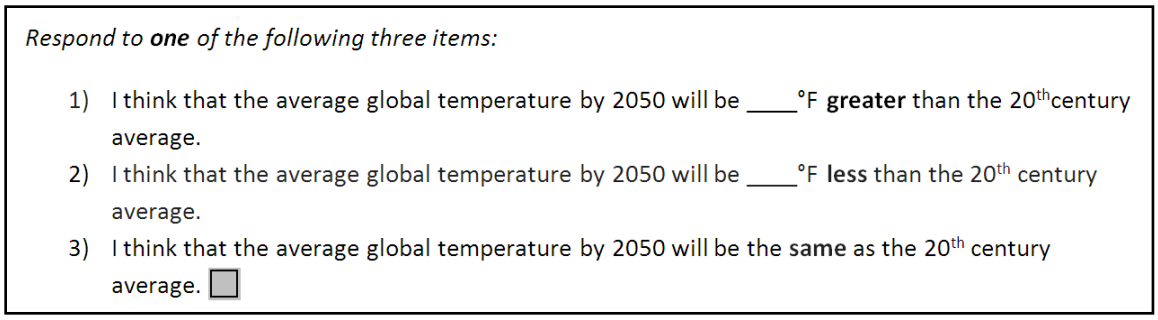
| | **M-Turk sample** | **National population (2012 Census data*)** |
|---|---|---|
| **Age** | Mean 33.42 years; median 30.0 years | Median 37.4 years |
| **Gender** | 56.7% female | 50.8% female |
| **Political party** | Republican 17.2%, Democrat 32.4%, Independent/other/unaffiliated 50.3% | Republican 24%, Democrat 29%, Independent 45% (2% unidentified)** |
| **Education** | 57.1% college educated + | 29.1% college educated + |

*U.S. Census Bureau (2014)
**Gallup (2014)

Stimulus and Measures. The questionnaire again presented respondents with two climate projections: a global average temperature increase of 3°F and a global average sea level rise of 6 inches over 20$^{th}$ century averages by 2050. (In this experiment, all participants were told the target year was 2050.) As before, after each projection, respondents completed five measures: their own parameter estimate, a lower-bound estimate, an upper-bound estimate, a rating of trust in scientists' estimate, and a rating of concern about the projection. Therefore, there were a total of 10 questions. The projections used in Experiment 2 were identical to those used in Experiment 1.

Experiment 2 explicitly asked participants to indicate if they thought future climate parameters would increase, decrease, or stay the same. They selected one of three options, filling in their estimates where applicable. For example, after reading the sea level projection, participants indicated if they believed the sea level would be greater than or less than the 20$^{th}$ century average, or if it would remain the same. They entered an estimate for either the first or second option, or checked a box marked "Yes" for the third option. See Figure 5.3.

Figure 5.3. Example parameter estimate questions in Experiment 2.




The questionnaire was programmed using the same online questionnaire tool as in Experiment 1, which was then linked to the Amazon Mechanical Turk website. The Internet link posted on the Mechanical Turk website was programmed to alternate between the two conditions in order to ensure an approximately equal number of respondents in each condition.

Design. The experiment had one between-participants independent variable, forecast format, with two levels: deterministic and uncertainty. In the deterministic format, single-value projections were shown for temperature and sea level; in the uncertainty condition, the projection values were accompanied by a range of possible values, which represented the upper and lower bounds of a 90% confidence interval. Half of the questionnaires showed single-value projections (deterministic condition), and the other half included estimates of uncertainty (uncertainty condition).

## Results

Hypotheses were the same as for Experiment 1: Inclusion of uncertainty estimates was predicted to increase participants' trust in and concern about the projections and to lead to higher and more accurate parameter estimates and estimates of parameter uncertainty. Again, before analyses, data from some respondents were removed. All of a respondent's data were excluded from subsequent analysis if one or more of the following occurred: the parameter estimate for

either temperature or sea level was extremely high (more than four times the projected increase); the lower- or upper-bound estimates for either parameter were extremely high (more than four times the projected upper bound of the predictive interval); the lower-bound estimate was greater than the upper-bound estimate; or the participant selected more than one option concerning future parameter value changes (e.g., temperature will increase by X°F *and* temperature will decrease by Y°F). This data cleaning process resulted in the removal of 209 participants from subsequent analysis, leaving 806 participants.

First, participants' parameter estimates were analyzed, both for temperature and sea level. Of 806 participants, 648 (80%) estimated a temperature increase, 17 (2%) estimated a temperature decrease, and 142 (18%) estimated no change; for sea level, 643 (80%) estimated an increase, 29 (4%) estimated a decrease, and 134 (17%) estimated no change. Independent-samples t-tests suggested that there was no difference between uncertainty and deterministic participants' parameter estimates. Lower- and upper-bound estimates, for both temperature and sea level, were not statistically different between conditions, although deterministic participants ($M$ = 3.63, $SD$ = 2.65) estimated greater temperature uncertainty (difference between lower- and upper-bound estimates) than uncertainty participants ($M$ = 3.22, $SD$ = 2.32) did, $t(786.47) = 2.31$, $p = .02$ (adjusted for unequal variances).

Next, trust and concern were analyzed. While there was no difference in concern between uncertainty and deterministic participants for either temperature or sea level, uncertainty participants trusted the projections more than single-value participants did. Independent-samples t-tests on trust for temperature and sea level revealed that uncertainty participants ($M$ = 3.66, $SD$ =1.30) trusted the temperature projections more than deterministic participants ($M$ = 3.47, $SD$ = 1.33) did, $t(804) = 2.05$, $p = .04$; and uncertainty participants ($M$ = 3.67, $SD$ = 1.34) trusted the

sea level projections more than deterministic participants ($M$ = 3.48, $SD$ = 1.33) did, $t(804)$ = 2.02, $p$ = .04.

**Discussion**

The results of Experiment 2, using a sample much more representative of the U.S. population than the sample used in Experiment 1, revealed a striking benefit of inclusion of uncertainty estimates. Participants who were shown climate projections that included uncertainty estimates trusted the projections more than did participants who were shown only single-value, deterministic projections. The increase in trust was observed for both temperature and sea level projections. This result bolsters the result of Experiment 1 on trust in climate projections and suggests that the benefit of inclusion of uncertainty estimates generalizes to the general population. Experiment 2 participants did not express any difference in concern between projection formats, although it should be noted the effect found in Experiment 1 was only marginally significant for temperature and not significant for sea level.

The effect of inclusion of the uncertainty estimate in Experiment 2 is particularly encouraging in that the sample used in Experiment 2 was substantially more like the general population than the sample used in Experiment 1 was. Whereas only 7% of Experiment 1 participants expressed (indirectly) doubt in climate change, approximately 20% of Experiment 2 participants expressed such doubt. While that is still short of the approximately 36% of Americans who polls suggest do not believe in climate change or are unsure (Leiserowitz et al., 2013), it is noteworthy that the effect of the uncertainty estimates was still demonstrated with a more representative, more skeptical population.

## General Discussion

Experiment 1 showed an effect of inclusion of uncertainty estimates on user trust, and Experiment 2, using a significantly more diverse and representative sample, showed the same effect. One can conclude that trust can be increased simply by including a range of values around a single-value climate projection and indicating the probability of observations falling within that range. This is consistent with some evidence that including uncertainty estimates in communications about environmental risk enhances users' perception of honesty in the source (Johnson & Slovic, 1995).

The effect might be driven by two different factors (or both). First, as was hypothesized, it might be that because people tend to have a good intuitive understanding of climate uncertainty, as was revealed in Experiment 1, a climate projection that includes a probabilistic expression of uncertainty seems more realistic and more complete, thereby making it seem more likely to be true. Second, it might be that for respondents with parameter estimates that did not match the single-value projections, the range of values in the uncertainty condition was more likely to include their estimates. Clearly, in both Experiment 1 and Experiment 2, there was variance in participants' parameter estimates, meaning that respondents gave projection estimates that did not always match the given projection. For these respondents, therefore, a projection that included a range of possible values that encompassed their estimates might well have seemed more trustworthy.

There were a few limitations of the present research, although some of them are inherent in this field of study. First, it is difficult to experimentally test perception of climate change projections because people already have specific knowledge about climate change (especially

people who visit weather blogs). Unlike study of perception of day-to-day weather forecasts, climate change is a hot topic with strong political undertones, and it is difficult, perhaps in some ways impossible, to isolate the cognition piece from complex constellations of belief. Furthermore, it is a dynamic issue, with public perception of climate change often shifting in the wake of highly publicized weather events (e.g., Weber, 2006). Second, the application of the results of my uncertainty manipulation (90% confidence interval) is limited by the fact that such estimates of uncertainty do not yet exist for climate projections. However, in the present experiment, the success of an interval expression, a common way to express uncertainty, nonetheless strongly suggests that communicating uncertainty in climate projections results in greater trust than communicating single-value projections. Furthermore, climate scientists are working to develop such estimates, so inclusion of the estimates in future climate projections is realistic. Finally, it is important to clarify that the manipulation discussed in this research is intended to provide better information to ultimately shift the beliefs of climate change skeptics, who now represent a diminishing portion of the population. The more difficult objective for future risk communication researchers is how to convince those who believe in climate change to be proactive in taking mitigating action. Still, convincing a sizable minority that climate change is occurring is a critically important piece, and the present findings are encouraging. Furthermore, increasing trust among those who already believe in climate change might improve the quality of their climate-related decisions, a topic that will be explored in the next chapter.

In summary, this research reveals positive effects on trust in climate projections by making the simple change of adding uncertainty estimates to them. The increase in trust in the projections might bring about more belief in climate change that ultimately translates to a more proactive population. Like with weather forecasts, climate projections that include uncertainty

estimates might well match users' intuition about uncertainty and lead them to have greater belief in the projections.

# Chapter 6: Making Climate-Related Decisions

## Background

The road salting experiments demonstrated that compared to conventional single-value forecasts, forecasts that included probabilistic uncertainty estimates resulted in better, more economically optimal decisions and higher trust in the forecasts. However, this improvement in performance did not apply to all the types of weather-related decisions tested. For example, in the rare weather event experiment, participants with probability estimates took significantly less precautionary action on the target trial, a weather event that was very unlikely but was nonetheless so extreme as to merit precautionary action. This one example suggests that inclusion of particular expressions of uncertainty might be more beneficial in some situations than in others; the means by which certain uncertainty expressions confer an advantage to decision makers might not apply in all situations.

One decision-making situation in which the effect of inclusion of uncertainty estimates is unclear is decision making with a large time horizon, such as making climate-related decisions. In the previous experiment's exploration of perception of climate change risk, participants expressed greater trust in climate projections that included uncertainty estimates. However, it is unclear if that trust would lead to enhanced decision making. The climate questionnaire was not a decision task. Evidence suggests that there is often disparity between people's judgments and decisions (Holt & Laury, 2005): A person might express preference for one course of action, but when actually faced with a decision consequence, they might reveal preference for another. If a person has to make a decision now about future climate, does it help or hurt to have a projection that includes estimates of climate uncertainty? If a builder were considering where to construct a

seaside vacation resort, would a climate forecast that included uncertainty estimates be helpful, perhaps allowing him or her to tailor planning decisions to his or her own risk tolerance? Would such forecasts be helpful to farmers faced with deciding which types of crops to plant, or would uncertainty estimates draw skepticism and distrust? It is not clear if the positive effects of inclusion of uncertainty estimates demonstrated earlier in decision tasks would be demonstrated in a far-future decision context.

Drought, generally "a deficiency in precipitation over an extended period" (NOAA, 2012a), is an example of an atmospheric phenomenon with potentially devastating consequences for which long-time-frame predictions can be calculated (Carbone & Dow, 2005; NOAA, 2012a). While there is debate among atmospheric and climate scientists about the degree to which the occurrence of drought will be affected by global climate change (Trenberth, 2012; Sheffield, 2012), drought will continue to be a serious problem with numerous consequences, including increased risk of forest fire and reduced agricultural productivity. In June 2012, for instance, a period of extreme heat and insufficient precipitation left 56% of the contiguous U.S. in drought conditions and contributed to the worst wildfire season in a decade, which burned 1.3 million acres in Colorado, Wyoming, and Utah (NOAA, 2012b). In January 2014, California suffered one of its most severe droughts in decades, causing wildfires to ignite, municipal water restrictions to be imposed, ski resorts to remain closed (due to lack of snow), and an estimated 200,000 acres of prime agriculture land to go unplanted in Fresno County (Onishi & Wollan, 2014). Many of these are largely unavoidable consequences of drought, but some of the effects could be mitigated. Farmers, for example, might be able to make better crop planting decisions if there were improved communication of drought projections, facilitating their decision of whether or not to plant crops more resistant to drought if the threat warranted doing so. Might including

probabilistic estimates of uncertainty in drought projections enhance decision making for decisions with distant outcomes like this?

It is unclear if the benefits to decision makers of having uncertainty estimates in weather forecasts would extend to decision makers faced with climate-related decisions. In this decision-making context, weather and climate are strikingly different in two ways. First, with weather-related decisions, decision makers receive prompt feedback, i.e., within a matter of hours or days; in contrast, with climate-related decisions, feedback might not be available for substantially longer periods, perhaps not for months, years, or even decades, quite possibly beyond the decision makers' lifetimes. Second, even if they are not consciously aware of it, most people have reasonably good intuition about weather, generated from a lifetime of day-to-day experience with predictions and outcomes (Morss et al., 2008). Climate projections, in contrast, are often vague and change over time, potentially as a result of the projections themselves (e.g., climate modeling includes anticipating what action societies will take to mitigate climate change), so it is difficult, if not impossible, to determine how good a climate projection is. These two factors suggest that climate-related decision making is different from weather-related decision making, and as such, uncertainty estimates might play a different role in each.

Similarly, two psychological mechanisms that might underlie the benefit of uncertainty estimates in weather-related decision making might not operate comparably in climate-related decision making. First, weather forecasts that include uncertainty estimates might seem less wrong when the predicted weather event does not occur because the forecast accounted explicitly for the probabilistic nature of the event. This quality might explain the increased trust in forecasts with uncertainty estimates and might make the forecasts more useful to decision makers. However, assessment of the forecasts' reliability requires feedback: Without outcome

information, decision makers might not realize the benefit of the uncertainty estimates' inclusion, or the additional probabilistic information – without any feedback – might simply make the forecasters themselves appear uncertain or incompetent (Frewer et al., 2003; Johnson & Slovic, 2005). Climate projections cannot be paired with any sort of readily observable outcome in a way comparable to weather forecast-observation pairs. Second, weather forecasts that include numeric uncertainty estimates might be deemed trustworthy because such forecasts match people's intuitions about weather uncertainty, intuitions that they have developed over a lifetime of exposure to forecasts and outcomes. Because people do not have experience (or have substantially less experience) with climate projections and outcomes, that source of trust – that the forecast expression matches intuition – would theoretically be absent. However, results from the climate questionnaire (Chapter 5) did suggest that people have an intuitive understanding of climate uncertainty: Participants estimated greater uncertainty as the projection year was farther in the future. It is possible, then, that even without direct experience of climate projections and outcomes, climate projections that include uncertainty estimates would more closely match people's understanding of climate uncertainty and lead to greater trust and better decisions.

There is some experimental evidence that people make better climate-related decisions with uncertainty estimates than without them. In one experiment (Patt, 2001), Zimbabwean farmers chose between two crops, maize and millet, and then spun a wheel with proportions representing varying probabilities of a wet year and a dry year. Choosing to plant maize when a spin resulted in "wet" earned participants $4, whereas an outcome of "dry" earned them $0. Choosing to plant millet earned them $2 for wet years and $1 for dry years. (These rewards matched the relative success of maize and millet in wet and dry years.) Results suggested that the rural farmers, with little to no formal math education, were able to make effective use of the

probabilistic information, making increasingly economically optimal decisions over several rounds of play. Uncertainty estimates had traditionally been withheld from the farmers on the belief that they would not understand them. But quite possibly because the probabilities used in the game concerned a domain very familiar to the farmers, i.e., it was not abstract, they were successful at making effective use of the information (Patt, 2001). However, as encouraging as these results were, it is not clear how well they generalize. First, the experiment included immediate decision feedback, which is often unavailable in real-world climate-related decision making. Second, participants were farmers, who likely had extremely well developed intuitions about climate variability and uncertainty, as their livelihoods were inextricably linked to those factors. The idea that an outcome – a wet or dry year – is inherently uncertain was likely well accepted by the farmers. While the climate questionnaire (Chapter 5) offered evidence that inclusion of uncertainty estimates in climate projections does indeed result in a modest increase in trust in the projection (among the general population), it is unclear if that slight increase in trust would result in people making better long-range decisions with projections that include uncertainty estimates. In summary, questions remain about the effect of inclusion of uncertainty estimates in a climate-related decision task.

Of critical interest in the present experiment is the role of decision feedback. Feedback is one of the two factors that fundamentally differentiates climate predictions from weather predictions. It is possible that the increased trust in weather forecasts that include uncertainty estimates, demonstrated in the road salting experiments, is due to feedback: In those experiments, trial after trial, participants saw outcomes, and perhaps participants with probabilistic uncertainty estimates perceived forecast-inconsistent outcomes as less inconsistent than did participants with only single-value forecasts, thereby increasing trust. In contrast, people

do not get immediate feedback when making climate-related decisions. Past research on the role of outcome feedback in decision making has been mixed (Kluger & DeNisi, 1996), although most evidence suggests that feedback is necessary for good decision making. However, the role of feedback has not been explored in the context of weather and climate, and it is unclear how the presence or absence of feedback interacts with uncertainty estimates or affects trust. Does feedback, inherently present in weather-related decisions and inherently absent in climate-related decisions, drive the "uncertainty estimate effect"?

To test this, I designed another decision task, roughly translating a realistic decision problem into a controlled setting. In the task, participants played the role of a consultant at an international agricultural consulting firm. They were shown a series of drought projections and had to advise hypothetical farmers on what type of crop – either drought resistant or not – to plant. Some participants received feedback on their choices, and others did not. Additionally, some participants were shown probabilistic uncertainty estimates with the drought projections, and others were shown only deterministic binary projections. The task was inspired by the crop-choice task used by Patt (2001).My drought experiment used a simplified representation of drought risk; in reality, there are numerous ways to define drought (e.g., NOAA, 2012a) and of course drought outcomes are not practically regarded as binary (drought versus no drought). Still, the experimental task captured important characteristics of the real-world problem: being faced with a difficult decision about a probabilistic future event with practical consequences, with a prediction about whether that event will occur or not. Also, the experiment employed a decision structure unlike the one used in the road salting task: Participants chose between two crops, one representing a sure gain and the other representing a mixed gamble, for which gain and loss outcomes are possible (Kahneman, 2011). This structure is representative of many real-

world decision scenarios in which there is both a cost and a potential profit. Most research on mixed gambles asks participants whether or not they would accept a given mixed gamble, allowing researchers to estimate loss functions (i.e., how much potential gain is necessary to offset a potential loss) (Hastie & Dawes, 2010). Very little research of which I am aware has asked participants to choose between a mixed gamble and a sure gain, as the present research does. In such a situation, participants' loss aversion would most likely lead to them being risk averse and tending to choose the sure gain because of the mere possibility of losing money in the mixed gamble (Payne, 2005).

I hypothesized that when feedback was provided immediately, trial by trial, participants with climate uncertainty estimates would make better decisions than participants who were shown only binary projections (drought or no drought), consistent with the earlier results from the road salting experiments. Without feedback, however, if feedback is indeed a critical part of the trust-inducing mechanism, I hypothesized that there would be no difference in performance between uncertainty participants and deterministic participants.

**Method**

The Drought Game was a computer-based task which participants completed in person at my computer lab. In the task, participants assumed the role of a consultant at an international agricultural consulting firm charged with helping farmers make sensible decisions about what crops to plant in the upcoming season. Of critical interest to the consultants was the risk of drought posed to farmers in different regions. Like in the road salting task, participants in the drought game were shown a series of drought projections, each pertaining to the drought risk in each farmer-client's region. On each trial, participants had to decide between two crops to

recommend to the farmer-client, Crop A or Crop B. Crop A cost $100 per acre to plant and would yield $300 per acre at harvest time, resulting in $200 profit, but only if non-drought conditions were observed. In the event of a drought, Crop A would die and would yield nothing, resulting in a net loss of $100 (the cost to plant the crop) on such a trial. Alternatively, Crop B cost $200 per acre to plant and would yield $300 per acre in any conditions, resulting in $100 profit. Therefore, participants had to choose between taking relatively costly but completely safe precautionary action (Crop B, with a sure profit of $100 per acre) or taking a gamble (Crop A, with a potential gain of $200 per acre but the risk of a net loss of $100 per acre). See Table 6.1. Unlike in the road salting task, the gamble was a mixed gamble between a gain and a loss, so participants chose between a sure gain (Crop B) and a mixed gamble (Crop A). Participants were given a starting budget of $1,000, which represented the profit per acre of their farmer-clients, and were instructed that they were to maximize their budgets. Additionally, after making their crop recommendation on each trial, participants gave ratings of trust in and concern about the drought projections on six-point scales. There were also overall ratings of trust and concern at the end of the task.

Table 6.1. Task cost structure: gains and losses associated with decisions to recommend Crop A (non-drought-resistant) or Crop B (drought-resistant).

| | **Outcome** | |
|---|---|---|
| | **Drought** | **No drought** |
| **Crop A** | Loss of $100 | Gain of $200 |
| **Crop B** | Gain of $100 | Gain of $100 |

Balancing the costs and potential profits of the crop options, the optimal course of action, given that the objective was to maximize the budget, was to choose Crop B (the risk-averse option) whenever the drought risk was greater than one-third. A 33% drought risk was the point

at which the expected payoff between Crop A and Crop B was equivalent. Whenever the drought risk was less than one-third, Crop A was the optimal choice. See Table 6.2.

Table 6.2. Expected profits associated with choosing Crop A and Crop B.

| | **Expected profit** | |
|---|---|---|
| **p(drought)** | *Crop A* | *Crop B* |
| 10% | $170 | $100 |
| 20% | $140 | $100 |
| 30% | $110 | $100 |
| 40% | $80 | $100 |
| 50% | $50 | $100 |
| 60% | $20 | $100 |

Of course, because the outcomes were probabilistic, on certain trials the optimal choice did not turn out to be the one with the higher payoff (e.g., appropriately choosing Crop A but experiencing a low-probability drought) , but in the long run, the optimal choices would pay best. See Table 6.3.

Table 6.3. Final balances and mean expected decision values associated with different choice strategies.

| **Decision strategy** | **Resulting final balance** | **Mean EV** |
|---|---|---|
| Always take gamble: Crop A on every trial | $6,000 | $109.25 |
| Always take sure gain: Crop B on every trial | $5,600 | $100.00 |
| Optimal strategy: Crop A when p(drought) < 33%, Crop B when p(drought) > 33% | $6,800 | $126.00 |

Participants. Three hundred fifty-one participants (43.0% female) participated in the experiment. Age ranged from 18 to 35 years ($M$ = 19.3 years). All participants were recruited from introductory psychology courses at the University of Washington and were awarded a small amount of academic credit for participating. Participants signed up voluntarily.

Procedure. Participants arrived in groups of 1 to 12 at pre-arranged times in a computer lab. After participants read a consent form and agreed to participate, the experimenter presented task instructions and a practice trial while participants followed along on their computers. Following an opportunity for questions, the task began. Participants completed the task individually. Upon completion of the task, participants answered some follow-up questions and were dismissed. All participants who showed up for the experiment were awarded academic extra credit. They did not receive any cash rewards.

Stimuli. There were a total of 46 trials in a fixed sequence. Drought conditions were observed on 14 of the trials (30.43%). Drought observations were roughly evenly spread throughout the 46 trials, with 7 in the first 23 trials and 7 in the second 23 trials. Drought projections, which were fictional, were restricted in range from 10% to 60% ($M$ = 30.43%), in multiples of 10%, so that they would be relatively rare. Actual drought projections range from 0 to 100% (see, for example, UC Irvine's GIDMaPS [Aghakouchak, Hao, & Nakhjiri, 2013] and Princeton University's Drought Monitoring and Hydrologic Forecasting [Luo & Pan, 2013]), but because of my interest in decision making about rare events, the projection set and cost structure of the task were geared toward relatively rare drought occurrences. The drought projections were reliably calibrated to the drought outcomes: Occurrence of drought matched the probabilistic prediction. For example, drought occurred on 10% of trials for which there was a projection of 10% likelihood of drought (1 drought out of 10 projections of 10% drought likelihood). For this reason, the overall proportion of drought trials exactly matched the mean likelihood of drought.

Design. There were two between-participants independent variables: projection format and feedback. The projection format variable had two levels: deterministic or probabilistic. Participants were shown either a binary drought projection implying determinism, i.e., either

there would or would not be drought, or they were given a probabilistic estimate of drought, e.g., "There is a 40% chance of drought in the farmer-client's region." The binary projection was based on the underlying probability of drought: If the probability were 33% or less, the projection indicated that drought would not occur, and if the probability were greater than 33%, the projection indicated that drought would occur. Note that this criterion, 33%, exactly matched the break-even point of the cost structure described earlier (i.e., choose Crop A if drought probability is less than 33%, Crop B if greater than 33%). The relationship between the location of the criterion and the break-even point is open to empirical investigation and could be systematically manipulated; here, the criterion matched the break-even point. The feedback variable also had two levels: feedback and no feedback. In the feedback condition, after making their crop recommendation, participants were told whether drought conditions were observed, and they were able to see their account balances throughout the task. In the no-feedback condition, participants were not told whether drought conditions were observed, and they were unable to see their balances until the very end of the task. Thus, the experiment had a 2 (projection format) x 2 (feedback) between-participants factorial design.

**Results**

It was hypothesized that the inclusion of drought uncertainty estimates would lead to better decisions and higher trust and concern, given that the results of the climate questionnaire extended to this decision task, but only when decision feedback was provided. Without feedback, it was predicted that there would be no difference between participants with and without uncertainty estimates. To address these predictions, mean expected decision value and ratings of trust and concern were analyzed. In all analyses, data from the first six trials were omitted. The first six trials were treated as practice trials, allowing participants time to become familiar with

the crop choice task, as the crop options were potentially difficult to understand immediately. The remaining 40 trials served as the experimental trials upon which the following analyses were conducted. Before analysis, data of questionable quality were removed. In the first several experimental sessions, a table presented in the task instructions, which showed the profits associated with Crops A and B when drought did and did not occur, contained a minor error. Out of an abundance of caution, the data of all participants who were shown the incorrect table were removed from subsequent analysis. There were 75 such participants, leaving 176 for analysis (42.6% female).

First, mean expected decision value was analyzed. For each participant on each trial, a value was calculated that reflected the long-run expected value of the decision. A decision to recommend Crop A was assigned the value of the probability of non-drought multiplied by the $300 revenue, less the $100 cost of the crop; a decision to recommend Crop B was assigned the value of $100 revenue (see Table 6.2). Those trial-by-trial values were then averaged so that each participant had a single mean expected decision value. Participants who were provided with probabilistic drought projections made significantly better decisions than participants who were provided with deterministic projections. A two-way ANOVA on mean expected decision value, with projection format (deterministic and probabilistic) and feedback (feedback and no feedback) as the independent variables, revealed a significant main effect for format, $F(1, 175) = 30.58$, $p < .01$. There was no effect of feedback, and the effect of feedback was approximately the same for both probabilistic and deterministic participants. The ANOVA revealed neither a significant main effect for feedback, $F(1, 175) = .07$, $p = .80$, nor a significant format by feedback interaction, $F(1, 175) = 1.17$, $p = .28$. See Table 6.4.

Table 6.4. Mean (standard deviation) expected decision value by feedback level.

| | **Deterministic** | **Probabilistic** | ***overall*** |
|---|---|---|---|
| **Feedback** | $118.63 ($6.91) | $122.51 ($4.22) | *$120.57 ($6.02)* |
| **No feedback** | $117.90 ($7.76) | $123.67 ($2.86) | *$120.79 ($6.50)* |
| ***overall*** | *$118.26 ($7.31)* | *$123.09 ($3.64)* | *$120.68 ($6.24)* |

As was discussed in previous chapters, mean expected decision value is a purer measure of performance than final balance in that it is an evaluation of decision quality irrespective of outcome, which is subject to chance. However, it can be noted that an analysis of final balance revealed an identical pattern of results: Participants who were provided with probabilistic drought projections performed significantly better at the task than participants who were provided with deterministic projections, $F(1, 175) = 5.45$, $p = .02$, and there was neither a significant effect of feedback nor a significant interaction. See Table 6.5.

Table 6.5. Mean (standard deviation) final balance by feedback level and projection format.

| | **Deterministic** | **Probabilistic** | ***overall*** |
|---|---|---|---|
| **Feedback** | $4,550.00 ($354.70) | $4,595.45 ($239.14) | *$4,572.73 ($301.62)* |
| **No feedback** | $4,547.72 ($384.88) | $4,713.64 ($171.98) | *$4,630.68 ($307.89)* |
| ***overall*** | *$4,548.86 ($367.97)* | *$4,654.55 ($215.44)* | *$4,601.71 ($305.28)* |

Next, trust and concern ratings were analyzed. Participants' ratings of trust in the projections did not differ significantly by projection format or by feedback level. Two-way ANOVAs on the mean of the trial-by-trial trust ratings and the single end-of-task trust rating, with projection format (deterministic and probabilistic) and feedback (feedback and no feedback) as the independent variables, revealed no significant effects. As for concern, participants who were shown deterministic projections expressed significantly greater concern than participants who were shown probabilistic projections. Two-way ANOVAs revealed significant effects for

the trial-by-trial concern measure, $F(1, 175) = 13.99$, $p < .01$, and for the final concern measure, $F(1, 175) = 6.11$, $p = .01$. There was no significant effect of feedback. See Tables 6.6 and 6.7.

Tables 6.6 and 6.7. Mean (standard deviation) trust and concern ratings (1-to-6 scale).

6.6. Trial-by-trial measure.

| | **Deterministic** | | **Probabilistic** | | ***overall*** | |
|---|---|---|---|---|---|---|
| | **Trust** | **Concern** | **Trust** | **Concern** | ***Trust*** | ***Concern*** |
| **Feedback** | 3.52 (.92) | 3.36 (.96) | 3.57 (.82) | 2.94 (.79) | *3.55 (.86)* | *3.15 (.90)* |
| **No feedback** | 3.56 (.71) | 3.32 (.57) | 3.74 (.78) | 2.91 (.51) | *3.65 (.75)* | *3.11 (.58)* |
| ***overall*** | *3.54 (.82)* | *3.34 (.79)* | *3.66 (.80)* | *2.92 (.67)* | *3.60 (.81)* | *3.13 (.75)* |

6.7. End-of-task measure.

| | **Deterministic** | | **Probabilistic** | | ***overall*** | |
|---|---|---|---|---|---|---|
| | **Trust** | **Concern** | **Trust** | **Concern** | ***Trust*** | ***Concern*** |
| **Feedback** | 3.59 (1.17) | 3.61 (1.13) | 3.84 (.86) | 3.30 (1.00) | *3.72 (1.03)* | *3.45 (1.07)* |
| **No feedback** | 3.55 (.85) | 3.66 (.96) | 3.66 (.83) | 3.25 (.78) | *3.60 (.84)* | *3.45 (.90)* |
| ***overall*** | *3.57 (1.02)* | *3.64 (1.04)* | *3.75 (.85)* | *3.27 (.89)* | *3.66 (.94)* | *3.45 (.99)* |

Concern was explored further by dividing the experimental trials into two groups: trials with low probability of drought (10-30%) and high probability of drought (40-60%), as one would expect concern about drought to be directly influenced by drought likelihood. Indeed, trial-by-trial concern was significantly higher on trials with high probability of drought. The repeated measures component of a mixed-model ANOVA, with drought probability (low and high) as the within-participants repeated measure and projection format and feedback as the between-participants variables, suggested that concern was significantly higher on high-probability trials than on low-probability trials, $F(1, 172) = 50.67$, $p < .01$. Importantly, participants with probabilistic projections differentiated low and high drought risk significantly more than participants with deterministic projections did: Whereas deterministic participants

expressed roughly the same concern for low- and high-probability drought projections, probabilistic participants expressed very low concern for low-probability projections and relatively high concern for high-probability projections. See Table 6.8. The mixed-model ANOVA revealed a significant interaction, $F(1, 172) = 13.07$, $p < .01$, suggesting that deterministic participants' concern ratings differed significantly less between the low- and high-probability trials compared to the difference in probabilistic participants' ratings between low- and high-probability trials. Feedback did not have any significant effect on concern ratings, nor did it interact significantly with drought probability. However, feedback seemed to influence the degree of concern participants expressed, depending on the format of the projections they were shown and the probability of drought. For deterministic participants, withholding feedback caused their concern to move more toward neutrality, i.e., higher for low-probability projections and lower for high-probability projections, whereas for probabilistic participants, the effect of withholding feedback was the opposite: They expressed even lower concern for low-probability projections and even higher concern for high-probability projections. The mixed-model ANOVA yielded a significant three-way interaction between the format, feedback, and drought probability terms, $F(1, 172) = 5.63$, $p = .02$, suggesting that the feedback manipulation had opposite effects on the format and drought probability variables, with concern increasing for low-probability trials and decreasing for high-probability trials for deterministic participants, but decreasing for low-probability trials and increasing for high-probability trials for probabilistic participants.

Table 6.8. Mean (standard deviation) trial-by-trial concern ratings at low and high probability of drought.

| | **Deterministic** | | **Probabilistic** | | ***overall*** | |
|---|---|---|---|---|---|---|
| | **Low prob** | **High prob** | **Low prob** | **High prob** | **Low prob** | **High prob** |
| **Feedback** | 3.15 (1.13) | 3.73 (1.13) | 2.66 (.87) | 3.47 (1.03) | *2.91 (1.04)* | *3.60 (1.08)* |
| **No feedback** | 3.28 (.96) | 3.39 (.99) | 2.46 (.72) | 3.74 (.83) | *2.87 (.94)* | *3.56 (.92)* |
| ***overall*** | *3.22 (1.05)* | *3.56 (1.07)* | *2.56 (.80)* | *3.60 (.94)* | *2.89 (.99)* | *3.58 (1.00)* |

Finally, decision strategy was explored further. Participants' choice of Crop B, the sure gain, was compared between trials with low probability of drought (10-30%) and high probability of drought (40-60%). Not surprisingly, participants selected Crop B more for high drought probability than low drought probability. (Note that because the decision was binary, not choosing Crop B meant that participants chose Crop A.) Deterministic participants chose Crop B more than probabilistic participants did overall, but, most importantly, participants with probabilistic projections differentiated low and high drought probability significantly more than participants with deterministic projections did, matching the results of the concern analysis above. That is, probabilistic participants appropriately chose Crop A on trials with low drought probability and chose Crop B on trials with high drought probability more than deterministic participants did. This was determined by performing a mixed-model ANOVA on proportion of trials on which participants chose Crop B, with drought probability (low and high) as the within-participants repeated measure and projection format (deterministic and probabilistic) and feedback (feedback and no feedback) as the between-participants variables, which suggested that participants selected Crop B significantly more on high-probability trials than on low-probability trials, $F(1, 172) = 1{,}169.41$, $p < .01$, and that deterministic participants selected Crop B more overall than probabilistic participants did, $F(1, 172) = 7.18$, $p < .01$. There was not a significant

main effect of feedback. However, the significant interaction between drought probability and projection format, $F(1, 172) = 9.34$, $p < .01$, suggested that probabilistic participants made more appropriate crop choices (Crop A for low probabilities, Crop B for high probabilities) than deterministic participants did. See Table 6.9.

Table 6.9. Proportion (standard deviation) of trials on which participants chose Crop B (sure gain) for trials with low and high drought probabilities by projection format.

| | **Deterministic** | | **Probabilistic** | | ***overall*** | |
|---|---|---|---|---|---|---|
| | **Low prob** | **High prob** | **Low prob** | **High prob** | **Low prob** | **High prob** |
| **Chose Crop B** | .19 (.23) | .84 (.21) | .07 (.13) | .85 (.18) | *.13 (.19)* | *.84 (.20)* |
| ***overall*** | *.51 (.14)* | | *.46 (.12)* | | *.49 (.14)* | |

These results can also be discussed in terms of decision errors. A participant committed a risk-averse error by selecting Crop B on trials with low drought probability; a risk-seeking error was selecting Crop A on trials with high drought probability. Participants made more risk-seeking errors overall, although the difference between risk-seeking and risk-averse error rates was not significantly different. Deterministic participants made more errors than probabilistic participants did. Interestingly, deterministic and probabilistic participants made risk-seeking errors at about the same rate, but they made risk-averse errors at different rates. Specifically, deterministic participants made more risk-averse errors than risk-seeking errors, whereas probabilistic participants made more risk-seeking errors than risk-averse errors. A mixed-model ANOVA on error rates, with error type (risk averse and risk seeking) as the within-participants repeated measure and projection format (deterministic and probabilistic) and feedback (feedback and no feedback) as the between-participants variables, suggested that participants overall made more risk-seeking errors than risk-averse errors, but not to a statistically significant degree, $F(1, 172) = 1.75$, $p = .19$. Deterministic participants made more errors overall than probabilistic

participants did, $F(1, 172) = 9.34$, $p < .01$. There was not a significant main effect of feedback. The significant interaction between error type and projection format, $F(1, 172) = 7.18$, $p < .01$, suggested that probabilistic participants made more risk-seeking than risk-averse errors, whereas deterministic participants made more risk-averse than risk-seeking errors. See Table 6.10. (Note that these final two analyses are mathematically identical to the earlier analyses about decision strategy and proportion of trials on which participants selected Crop B. The present analyses simply frame the same data in terms of decision errors.)

Table 6.10. Proportion (standard deviation) of trials on which participants committed errors by projection format.

| | **Deterministic** | | **Probabilistic** | | ***overall*** | |
|---|---|---|---|---|---|---|
| | **Risk averse** | **Risk seeking** | **Risk averse** | **Risk seeking** | **Risk averse** | **Risk seeking** |
| | .19 (.23) | .16 (.21) | .07 (.13) | .16 (.18) | *.13 (.19)* | *.16 (.20)* |
| ***overall*** | *.18 (.16)* | | *.11 (.10)* | | *.15 (.14)* | |

**Discussion**

The results of the drought experiment offer yet more evidence of the benefit conferred to decision makers who are given numeric uncertainty estimates with which to make their decisions. Results were largely consistent with the experiments reported earlier: Participants who were shown probabilistic drought projections made significantly better decisions than did participants who were shown deterministic projections, as measured by mean expected decision quality, and that superior decision making led to better task performance overall, as measured by final balance. These results support the hypotheses.

It was predicted that participants would be biased toward committing risk-averse errors, selecting the sure gain of Crop B more than was economically optimal. Results suggested that

participants were not biased toward risk aversion. However, that might have been because of the influence of the uncertainty estimate. The significant error by projection format interaction suggested that deterministic participants indeed made more risk-averse errors than risk-seeking errors, consistent with the prediction that participants would be risk averse when choosing between a mixed gamble and a sure gain. However, inclusion of uncertainty estimates in the drought projections largely eliminated that bias: Probabilistic participants made fewer risk-averse errors than risk-seeking errors, and moreover they made only about one-third as many risk-averse errors as deterministic participants did. This strongly suggests that in situations in which people might be inclined to make errors of risk aversion, including a probability estimate combats the effect.

Participants' ratings of trust in the projections did not differ significantly by projection format or by feedback level, although mean differences by format were in the predicted direction: Probabilistic participants gave higher trust ratings than deterministic participants did. Deterministic participants gave higher ratings of concern than did probabilistic participants, contrary to the hypothesis. However, this unpredicted concern effect was likely largely driven by the very low concern for low-probability droughts expressed by probabilistic participants. Because of the focus on relatively rare events, the majority of trials in the task had a low probability of drought. (The highest drought probability was only 60%.) Most trials – 31 out of 40 experimental trials, or 78% – had a drought probability of less than half, and those participants who were made explicitly aware of the less-than-even-odds drought likelihood might therefore have expressed relatively little concern compared to participants who were simply told drought was or was not projected. For deterministic participants, a projection of "drought is projected" – which occurred on 14 out of 40 trials, or 35% – might have been interpreted to

mean that drought was 100% likely, thereby generating greater concern overall. In comparison, that means that probabilistic participants were shown that there were merely even odds or slightly-better-than-even odds of drought on only 22% of trials, whereas deterministic participants were shown that drought was projected – potentially meaning that it was *certain* to occur – on 35% of trials. Therefore, in retrospect, it is not surprising that deterministic participants gave higher concern ratings, both trial by trial and at the end of the task.

Comparing concern ratings on low- and high-probability trials demonstrated an extremely important attribute of uncertainty estimates: They provide information that is more specific than mere deterministic predictions. Probabilistic participants in the present experiment revealed a greater differentiation of concern between low- and high-probability projections. It is critical to note that probabilistic participants therefore expressed more appropriate amounts of concern than did deterministic participants. The overall higher concern rating of the deterministic participants discussed above is slightly misleading. More concern is not necessarily better. The low-probability versus high-probability analysis showed that probabilistic participants were less concerned about low-probability droughts than deterministic participants were and more concerned about high-probability droughts than determinist participants were, ultimately leading to fewer risk-averse errors and better task performance (e.g., higher mean expected decision value).

The significant three-way interaction was also an important finding. In reality, when shown climate projections, people will not receive immediate feedback. The "no feedback" level of the feedback variable was designed to capture that. For probabilistic participants, not being shown feedback enhanced their differentiation in concern between low- and high-probability projections: Without feedback, they felt less concern for low-probability droughts and more

concern for high-probability droughts, compared to the concern they expressed when shown feedback. This differentiation in concern is appropriate, both within the context of the drought experiment (in which low-probability projections did not warrant taking the precautionary action of choosing Crop B) and in real life. Deterministic participants, on the other hand, demonstrated the exact opposite tendency, expressing more concern for low-probability droughts and less concern for high-probability droughts, compared to the concern they expressed when shown feedback. Clearly, the pattern of behavior demonstrated by probabilistic participants is superior to that demonstrated by deterministic participants when considering the lack of feedback of climate projections in real life.

It should be noted that the deterministic projections in the drought experiment – a binary projection of drought versus no drought – were different from those in the experiments reported earlier. In this experiment, the deterministic projection was similar to the binary advice of the Decision Support Aid in the decision advice experiment: It unambiguously indicated what the participant ought to do on each trial. The deterministic projection used the cost-revenue-based break-even point (drought probability of 33%) as its criterion, so when the probability of drought was greater than 33%, drought was projected to occur, and when it was less than 33%, drought was projected not to occur. Basing decisions on the deterministic projection, therefore, would result in economically optimal performance. This was not the case in the decision advice experiment, in which basing decisions on only the single-value temperature estimate (i.e., salt if the forecast were 32°F or less, withhold salt if it were 33°F or more) would not lead to optimal performance. In other words, in the drought experiment, theoretically the deterministic projections were particularly useful. Therefore, that participants given numeric probability

estimates *still* outperformed deterministic participants is strong evidence for the value of inclusion of uncertainty estimates.

Finally, the lack of effect of the feedback manipulation is highly noteworthy. First, it is not consistent with much past research suggesting that feedback is necessary for good decision making (e.g., Einhorn & Hogarth, 1981). Absence of decision feedback did not adversely affect participants' task performance. This surprising result needs to be replicated. Second, the lack of a feedback effect gives clues about the mechanisms underlying the beneficial effect of inclusion of uncertainty estimates. One possible mechanism to explain why uncertainty estimates are helpful to decision makers is that projections that include uncertainty estimates might seem less wrong when the predicted event (in this case, drought) does not occur because the projection accounts explicitly for the probabilistic nature of the event. Neither an occurrence nor a non-occurrence of the event contradicts a probabilistic prediction. However, the present results suggest that getting feedback did not make any difference. Trust was just as high when feedback was withheld as when it was given. This then suggests that the more likely explanation of how uncertainty estimates benefit decision makers is that projections that include numeric uncertainty estimates provide better information that matches people's intuitions about future uncertainty. Even without direct repeated exposure to predictions and outcomes, in general people correctly intuit uncertainty in the future climate, consistent with the results of the climate questionnaire experiment. Although probabilistic participants in the present experiment did not express greater trust in the projections than deterministic participants did, inclusion of uncertainty estimates did not reduce trust, so inclusion of uncertainty estimates in climate projections seems well warranted.

Overall, this experiment offers compelling converging evidence for the benefit of inclusion of numeric uncertainty estimates. Participants performed an experimental task concerning a farther-future event than had been previously tested, and with or without being provided decision feedback, participants with probability estimates did better than those without it. Participants with the uncertainty estimates had more appropriately placed concern, and although they did not express greater trust in the projections, inclusion of uncertainty estimates did not diminish their trust. These results strongly suggest that uncertainty estimates are beneficial in an ever-widening array of scenarios and that their inclusion in climate projections would benefit farmers and anyone else faced with a climate-related decision.

## Chapter 7: General Discussion

The goal of this body of research was to expand upon recent research concerning the benefits of inclusion of probabilistic uncertainty estimates in weather forecasts and explore the extent of those benefits. In summary, this set of experiments generated compelling evidence for numerous benefits of inclusion of uncertainty estimates, particularly numeric probability, and for conditions in which expressions of weather uncertainty other than numeric probability are superior.

First, results revealed that uncertainty estimates reduce the negative impact on decision making of high-error forecasts. This is a critical finding, as arguably some of the most important forecasts are ones concerning severe weather or storms that are made long enough in advance to give people time to pursue precautionary action. At such long lead times, forecast error can be high, but including uncertainty estimates might, over time, preserve public trust in long-lead-time weather warnings and lead to better decisions in response to potential weather threats. Results suggested that uncertainty estimates combat the adverse effects of false alarms, as well. False alarms negatively impact decision makers' decision quality and trust in forecast sources, but the "uncertainty estimate effect" is stronger than the effect of reducing false alarms: Adding uncertainty estimates to forecasts is a more effective means to promote compliance with weather warnings than adjusting the false alarm rate of decision advice.

The experiments demonstrated that inclusion of uncertainty estimates is overall preferable to providing explicit decision advice as a means to promote good decision quality, user trust, and compliance with weather warnings. However, at low probabilities of an event at which precautionary action is warranted or economically optimal, combining decision advice and

uncertainty estimates leads to the best decisions overall. Doing so reduces decision errors, notably the risk-seeking errors people might make when a low numeric probability alone makes taking precautionary action seem counterintuitive. An effective way to express the risk of rare events is with an odds ratio. Results suggested that compared to probability estimates, the odds ratio is a more powerful way to express rare but important weather events. Expressing the increase in likelihood of rare events by using odds ratios increases precautionary action for events that warrant such action, but also for events that do not warrant precautionary action (i.e., providing relative risk estimates does not improve people's ability to discriminate between appropriate and inappropriate precautionary action), so great care must be taken when using this expression of uncertainty. Numeric probability estimates, therefore, are an effective way to communicate weather uncertainty and lead to better decisions overall, but decision advice or a relative risk expression communicates low-probability or rare risks more successfully.

Finally, the effect on trust of communicating weather uncertainty extends to the communication of climate change uncertainty. Including climate uncertainty estimates (over conventional single-value estimates) leads to a modest increase in trust for projections concerning two important climate parameters. This is a significant result, as there is considerable effort and attention lavished on promoting better decision making about climate change, and adding uncertainty estimates to climate projections is a simple and straightforward intervention. Additionally, decision makers faced with far-future uncertainty can benefit from uncertainty estimates. Including a probabilistic estimate of drought leads to superior decision making and more appropriately placed concern about drought. Trust in projections with explicit drought probabilities is not greater than trust in deterministic projections, but importantly, it is not less, either. Feedback does not interact with inclusion of uncertainty estimates, suggesting that the

usefulness of uncertainty estimates is conferred in their matching people's intuitions about weather and climate uncertainty.

Taken together, these results are very encouraging in two ways. First, they strongly suggest that non-expert members of the public can make effective use of and benefit from probabilistic information. With past research (Nadav-Greenberg & Joslyn, 2009; Roulston et al., 2006) as a starting point, the present collection of experiments shows that for a variety of weather-related decision problems, uncertainty estimates – expressed in different ways depending on the particular problem – reliably improve decision quality. They also generally increase trust, most likely by matching people's intuitions about weather uncertainty, thereby making the forecasts seem more complete.

Second, in a more practical way, the results suggest that decision quality in a number of highly consequential real-life situations could be improved with the simple addition of uncertainty estimates to forecasts. In the experiments reported here, the forecasted events (e.g., freezing temperature, drought) occurred relatively rarely and as such might have led participants to underestimate their risk, leading to risk-seeking decisions (Hertwig et al., 2004). Including uncertainty estimates in the forecasts and projections reduced participants' tendency to seek risk, thereby improving decision quality. Furthermore, beyond risk seeking due to this underestimation of risk, the decision tasks used in the experiments featured cost structures associated with decision biases: the road salting task was framed in terms of losses, in which people tend to make risk-seeking errors, and the drought task was framed mostly in terms of gains, in which people tend to make risk-averse errors. The results of the experiments reported here provide compelling evidence that including uncertainty estimates improves decision making in both types of frames.

The results are also encouraging in that they could be conservative. The experiments revealed improved decision quality with uncertainty estimates even though participants had essentially no experience with using such information for the decision tasks at hand. Participants' use of freeze probability estimates to make road salting decisions, for example, was likely a completely novel experience for them. With repeated exposure over a long period of time, people might develop richer intuition about numeric uncertainty estimates, augmenting their pre-existing intuition about weather uncertainty in general. It has been noted that even if people do not fully understand the meaning of certain weather processes or concepts, such as barometric air pressure, over time they develop a practically useful understanding of how the processes work and what the concepts mean (Morss et al., 2008). An example of this is probability of precipitation, which has been communicated to the public for decades. Even though evidence suggests people often misinterpret what that probability means (e.g., Joslyn, Nadav-Greenberg, & Nichols, 2009), it is now an indispensible part of weather forecasts that people value having (Morss et al., 2008). The results reported in the present dissertation research demonstrate that people benefit from such uncertainty expressions as freeze probability, with no apparent direct misunderstanding, suggesting that there really is little reason not to include them in public weather forecasts.

Overall, this research, and hopefully more like it in the future, provides evidence for the way humans respond to different expressions of uncertainty, evidence that ideally could inform how weather uncertainty is communicated to the public. This type of research is critical as high-quality information is increasingly available. Without communicating it effectively, its value cannot be realized.

However, it must be noted that communication of probabilistic uncertainty estimates should be carefully considered. Inclusion of such information can lead to unintended problems, such as the overestimation of risk of the lowest-probability target event demonstrated by odds ratio participants in the rare weather event experiment. In one experiment in which participants were presented with hypothetical hurricane threats, participants' perception of hurricane risk to them was based partly on an assessment of how great the risk was to others in nearby areas (Baker, 1995), suggesting that even if one would typically pursue precautionary action in response to a particular level of weather-related risk, seeing that others are in greater risk might make one underestimate his or her own risk. Additionally, research suggests that people often read additional information into uncertainty estimates, such as incorporating base rate probability into event-specific probability estimates, e.g., evaluating a 20% chance of rain differently for Seattle and for Las Vegas (Windshitl & Weber, 1999), and attempting to decode what is sometimes perceived to be forecasters' crossing of likelihood and severity, potentially resulting in underestimation of risk (Patt & Shrag, 2003). The bottom line is that although some uncertainty expressions and communication formats are better suited for some circumstances than others, interpretation errors can still occur and communicators should be highly aware of this.

Still, the results demonstrate that people make better decisions with uncertainty estimates than without them in a number of weather-related decision settings. When making weather-related decisions in both the real world and in the laboratory setting of the present experiments, people face two complicating factors that are essentially inherent in weather forecasts: prediction error and false alarms. Both are born of weather uncertainty. These factors were systematically manipulated in two experiments reported here, showing that increasing these factors (i.e.,

increasing the forecast error and increasing the false alarm rate) leads to a clear decline in decision quality and trust. The factors were not directly tested in the other decision-making experiments, but forecast error was always present and many participants likely experienced taking precautionary action when it turned out to be unnecessary. The experiments closely match the real-world experience in this respect, as forecast error and false alarms are ubiquitous in real weather forecasts and emergencies. Even if interpretation errors can occur when uncertainty estimates are included in weather forecasts and warnings, the evidence presented here suggests that in the face of forecast error and false alarms, people are better off having such estimates than not having them.

Much of this research is novel. The topic of weather-related decision making is widely explored in the field, but there is extremely little experimental evidence of the factors affecting weather-related decision making. Very little past research has systematically manipulated these factors in a controlled experimental setting. The research reported here represents a substantial contribution to our understanding of how numerous factors, such as forecast error, expression of weather uncertainty, false alarms, and decision feedback, affect the decisions people make when faced with weather uncertainty. This research also contributes to our understanding of decision making with mixed gambles and the relationship between prediction error and trust. Further experimental evidence and replication of my results is necessary, but the evidence from this set of experiments is a significant start. The consistent patterns of results reported above, generated from several different experimental paradigms, constitutes strong converging evidence about the way people understand weather uncertainty, how they make decisions when faced with it, and how their decision making improves in a variety of settings when forecasts and projections include quantified estimates of uncertainty.

The present research also demonstrated several psychological phenomena which have been explored more extensively in basic research. The predominant decision error demonstrated in the present research was seeking more risk than was economically optimal. Risk seeking was demonstrated in the experimental tasks by not taking sufficient precautionary action. This was likely due to the fact that the events being forecasted in the experiments, like freezing temperature and drought, occurred relatively rarely in the trial sequences. Research suggests that in such settings, in which people complete multiple trials of a task and experience the low incidence of low-probability events (as opposed to having the probabilities merely described to them), people tend to underweight the likelihood of the events occurring (Hertwig et al., 2004; Erev & Barron, 2005). The results of the present experiments are consistent with this: Overall, participants tended to seek more risk than was economically optimal. This risk-seeking tendency was particularly pronounced in the road salting experiments, in which participants had to choose between a sure cost and a gamble with a potentially greater loss. As predicted by prospect theory (Kahneman & Tversky, 1979), with choices framed in terms of losses, participants tended to make substantially more risk-seeking than risk-averse errors. This is consistent with the real-world phenomenon of people tending not to take adequate precautionary action when faced with weather-related risks and instead taking a gamble, as presumably the chance of a substantial loss is less aversive than a sure loss. Although risk aversion at low probabilities has been observed in tasks framed in terms of losses (Tversky & Kahneman, 1992), the probabilities used in the present loss-framed experiments might have been sufficiently high to elicit risk seeking, or perhaps the rareness of the forecasted events (freezing temperature, drought) in the present experiments led to probability underweighting that overcame bias toward risk aversion. The risk-seeking tendency was less pronounced in the drought experiment, in which participants had to

choose between a sure gain and a mixed gamble offering a possible greater gain or a possible loss. This reduced risk-seeking tendency was likely due to framing, which was mostly in terms of gains: Past research suggests participants would demonstrate a risk-averse bias in this situation (Birnbaum & Bahra, 2007; Payne, 2005). In this case, risk-seeking and risk-averse tendencies (due to rare event occurrence and gains frame, respectively) might largely have canceled each other out. Indeed, there was not a significant difference between participants' rates of risk-seeking and risk-averse errors overall in the drought experiment, but results suggested that participants with conventional single-value projections did indeed make more risk-averse than risk-seeking errors. The inclusion of an uncertainty estimate effectively neutralized the risk-aversion main effect.

Some results were inconsistent with past research, however. In the drought experiment, decision feedback appeared to have no effect on participants' task performance. It did not affect their trust in the drought projections, and it did not interact with the inclusion of uncertainty estimates. These results are largely inconsistent with past research suggesting that outcome feedback is essential for learning and for good task performance (Einhorn & Hogarth, 1981; Kahneman & Klein, 2009; Shanteau, 1992). The results were also surprising because feedback was hypothesized to play a central role in the improvement of decision making when uncertainty estimates are included. Decision feedback must be present for a decision maker to evaluate how well an outcome matches a prediction. It was hypothesized that prediction-outcome mismatches would not be perceived as incorrect if the prediction included an uncertainty estimate, and that that perception explains (or partly explains) why participants typically give higher ratings of trust to weather forecasts that include uncertainty estimates. In other words, it was expected that among participants who were provided with uncertainty estimates, those who were given

feedback would give higher ratings of trust and make better decisions than participants without uncertainty estimates, whereas those who were not given feedback would perform no better than control (single-value) participants. The lack of feedback effects in the drought experiment suggest prediction-outcome evaluation does not drive the "uncertainty estimate effect." The unexpected results suggest that the drought experiment must be replicated in order to confidently conclude that decision feedback is indeed unrelated to good performance at this task and does not interact with inclusion of uncertainty estimates.

Classic frequency format effects were not demonstrated, either. The experimental results were perfectly neutral concerning the ease of understanding frequency versus probability expressions. Participants in experiments in which those forecast formats were tested demonstrated comparable decision quality and indicated comparable levels of trust. As was noted earlier, it might be that the frequency format offers no particular advantage over probability in the domain of weather, because people recognize that a percent could theoretically refer to the climatological record. It could also be that the advantage for frequency formats arises in situations in which risks need to be compared or in which calculations are necessary, unlike in the sorts of binary choice tasks tested in the present research. However, even if that were the case, many real-world weather-related decisions are indeed binary (e.g., deciding between staying home or evacuating), so in this practical context the difference in psychological mechanisms underlying understanding of probability and frequency might be irrelevant. Additionally, some research suggests that probabilities and frequencies are understood roughly equally (Cuite et al., 2008).

It should be noted that in my frequency manipulation, the denominator was always 100, e.g., "on 20 out of 100 days like this" and "on 45 out of 100 days like this." Frequency effects

demonstrated in past research might be due partly to denominator neglect (Kahneman, 2011), meaning that focus is directed primarily toward the numerator (Lipkus, 2007; Yamagishi, 2007). I could have systematically varied the denominator in a more complete test of frequency formats, e.g., "on 200 out of 1,000 days like this" in addition to "on 20 out of 100 days like this." Perhaps such a manipulation would have led to frequency effects, with "200 out of 1,000" perceived as expressing greater likelihood than "20 out of 100" or "20%." As tested, my frequency and percent probability formats were directly numerically comparable, as both used a denominator of 100, so even if the denominator were neglected, there would have been no difference between formats. Still, it is interesting that the perception of countable instances of events (e.g., "20 days") was perceived no differently from the more abstract probability (20% chance), as one might predict based on some past research (Slovic et al., 2004).

The positive effect of the inclusion of uncertainty estimates on decision quality could perhaps be explained through the lens of the two-systems perspective. One quality of System 1 is that it attempts to construct meaning from incoming information by first "making it true," creating a coherent story with a cause and an effect (Kahneman, 2011). When reading single-value, deterministic forecasts, therefore, perhaps understanding the forecast information was merely a System 1 process: Participants read it, interpreted it as a true statement (e.g., the nighttime low temperature will be 33°F) and then made decisions accordingly. Such a process would immediately result in being faced with the predictive error of the forecasts; thereafter, perhaps participants ignored the forecasts or relied on some other intuition to make decisions. Clearly, those participants revealed a lack of trust in the forecasts. In contrast, forecasts that included numeric probability estimates were perhaps sufficiently difficult to understand that they required engagement of System 2 processes, which then allowed for more deliberate and

analytical cognition and ultimately led to better decisions. In a sense, the uncertainty information might have been inducing cognitive strain, thereby activating System 2 processing. Deterministic forecasts, on the other hand, might have been processed with relative fluency, requiring no more than automatic, intuitive System 1 processes, until recognition of the error caused participants to lose faith in the predictions.

Of course, the effect of uncertainty estimates might have been more simply due to the fact that numeric probability information is of finer granularity and therefore provides richer input on which to base decisions. Consider the temperature range of only 6°F in the single-value forecasts in the original road salting experiment (32°F to 37°F, inclusive), compared to the range in freeze probability of 42 percentage points (10% to 51%, inclusive). The latter offers significantly more precise information. Still, that does not discount the possibility that System 2 forces were engaged more among uncertainty participants than among deterministic participants. The most important point is that including uncertainty estimates could potentially circumvent many of the heuristic errors discussed earlier. Affective factors, recency, and availability, for example, are all independent of the actual probability of a future weather event. For example, if there's a 20% chance of a flood, the probability of the flood is 20% regardless if a nearby resident has vivid memory of another recent flood. Surely the recency of the other flood and the vividness of the memory might enhance the resident's perception of the predicted flood's likelihood, but communicating the 20% probability explicitly might lessen the influence of those factors. The effect of recent experience with an adverse weather event was not directly tested in the present body of experiments, but the improved decision making with uncertainty estimates that was demonstrated here suggests that uncertainty estimates would reduce the effects of recent experience more than would conventional single-value forecasts. Uncertainty estimates make for

a richer forecast that likely requires more deliberate, analytical processing, thereby reducing the influence of more affect-driven responses to recent experience. For example, in the road salting task, recent experience with a miss (not applying treatment and a freezing temperature occurs) might lead to unnecessarily risk-averse decisions in subsequent forecasts because of the mere averseness of the loss. Inclusion of uncertainty estimates in the forecast might engender more analytical evaluation of subsequent forecasts and lead to better decision making.

A limitation of the present research is that the laboratory tasks were drastic simplifications of the actual tasks they theoretically represented. For instance, in the road salting experiments, an actual road-salting decision would include numerous other factors besides the temperature (and freeze probability) forecast (Roulston et al., 2006), not the least of which is the precipitation forecast, which is wholly left out of the experimental paradigm. Additionally, the consequences faced by real-world decision makers might not be represented adequately by incentives to participants of less than $10. The possibility of payment for good performance was intended to represent a "real consequence," but it is quite possible that such small rewards in the experiments did not represent the weightiness of the real-world decision-consequence relationship. Whereas the real consequence to participants in the road salting tasks was the possibility of winning cash, the real consequence to decision makers in many weather emergencies is loss of life. The limitation of the present research, therefore, is simply the degree to which the results can be generalized to real-world situations. Still, the results reported here replicate patterns of behavior demonstrated in actual weather emergencies (e.g., not taking precautionary action when it is advised) and provide a clear line of evidence of the influence of uncertainty estimates on decision making. There is an unambiguous effect, an effect that might be demonstrated outside the context of the laboratory and even in other domains. If there were no

relationship between inclusion of uncertainty estimates and decision making, none would have been demonstrated in the experiments.

An interesting extension of the present research would be to explore certain variables that theoretically influence the way people make decisions when faced with uncertainty. For example, future experiments could systematically explore the role of precautionary action versus inaction. In the road salting experiments, risk was averted by taking the precautionary action of applying treatment to the roads, i.e., a decision to act. However, in other situations, risk is averted by taking the precautionary action of withholding some type of action, such as staying home upon hearing a warning about a snow storm, i.e., a decision to *not* act. It is unclear how the results of experiments in which participants exercise action-oriented precaution generalize to situations in which people exercise inaction-oriented precaution. That variable is hardly irrelevant; indeed, ample research suggests that when faced with uncertain outcomes, people are biased toward inaction (omission bias: Ritov & Baron, 1995), and the role of uncertainty estimates in inaction-oriented situations is unclear.

For another example, future experiments could systematically explore the self-other dimension of decision making. In both the road salting experiments and the drought experiment, participants made decisions for others, and the results of those experiments were generalized to how people make decisions for themselves. However, a small body of research suggests that despite numerous inconsistent findings across experiments, the role of the decision maker (i.e., as deciding for self or other) to some extent determines the way decisions are made, notably that one tends to believe others are less affected by emotion than they actually are, resulting in making more conservative or risk-neutral decisions for others than for oneself (e.g., Ziegler & Tunney, 2012; Polman, 2010; Garcia-Retamero & Galesic, 2012). Future research, therefore,

should systematically explore these variables to provide further clarity and nuance to this complex topic.

In conclusion, the research reported here offers an optimistic view of our ability to make decisions when faced with uncertainty. It provides a clear line of evidence that adding numeric uncertainty estimates to weather forecasts and climate projections results in better decisions, higher trust, and greater compliance with warnings. The consistent pattern of results generated in multiple experimental settings offers converging evidence of a psychological effect. While the real-world situations this set of experiments was designed to replicate are surely more complex than the controlled experiments themselves, this converging evidence strongly suggests that the findings of these experiments should be tested in real-world settings. Ultimately, emergency managers could communicate weather-related risk more clearly, people could make better decisions, and lives could even be saved. Furthermore, the basic pattern of results demonstrated here might be applicable in other domains, such as medicine and finance. Testing might reveal that the improvements in decision making with uncertainty estimates demonstrated here extend well beyond weather and climate.

## References


Aghakouchak, A., Hao, Z., & Nakhjiri, N. (2013). Global integrated drought monitoring and prediction system. Accessed online on February 2, 2014, at http://drought.eng.uci.edu/GIDMaPS_Doc.pdf.

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association, 13* (6), 608-618.

Ando, M., Wakai, K., Seki, N., Tamakoshi, A., Suzuki, K., Ito, Y., Nishino, Y., Kondo, T., Watanabe, Y., Ozasa, K., & Ohno, Y. (2003). Attributable and absolute risk of lung cancer death by smoking status: Findings from the Japan Collaborative Cohort Study. *International journal of cancer, 105* (2), 249-254.

Ariely, D. (2010). *Predictably Irrational*. New York: Harper Perennial.

Bach, D. R., Hulme, O., Penny, W. D., & Dolan, R. J. (2011). The known unknowns: Neural representation of second-order uncertainty, and ambiguity. *The Journal of Neuroscience, 31* (13), 4811-4820.

Baddeley, A. (1992). Working memory. *Science, 255*, 556-559.

Baker, E. J. (1991). Hurricane evacuation behavior. *International Journal of Mass Emergencies and Disasters, 9* (2), 287-310.

Baker, E. J. (1995). Public response to hurricane probability forecasts. *The Professional Geographer*, 47, 137-147.

Birnbaum, M. H., & Bahra, J. P. (2007). Gain-loss separability and coalescing in risky decision making. *Management Science, 53* (6), 1016-1028.

Blendon, R. J. (2008). *High-risk area hurricane survey*. Accessed online on April 2, 2012, at http://sphweb.sph.harvard.edu/news/press-releases/files/Hurricane_2008_Total_Release_Topline.doc.

Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. *Ergonomics*, *43* (9), 1283-1300.

Bliss, J. P., & Fallon, C. K. (2006). Active warnings: False alarms. In M. Wogalter (ed.), *Handbook of Warnings* (Chapter 17, 231-242). Mahwah, NJ: Lawrence Erlbaum.

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, *38* (11), 2300-2312.

Bostrom, A., & Lofstedt, R. E. (2003). Communicating risk: Wireless and hardwired. *Risk Analysis, 23* (2), 241-248.

Boykoff, M. T. (2009). We speak for the trees: Media reporting on the environment. *Annual Review of Environment and Resources*, *34*, 431-457.

Brase, G. L. (2002). Which statistical formats facilitate what decisions? The perception and influence of different statistical information formats. *Journal of Behavioral Decision Making, 15* (5), 381-401.

Breznitz, S. (1985). False alarms: Their effects on fear and adjustment. *Issues in Mental Health Nursing*, *7* (1-4), 335-348.

Brooks, H. E., Witt, A., & Eilts, M. D. (1997). Verification of public weather forecasts available via the media. *Bulletin of the American Meteorological Society, 78* (10), 2167-2177.

Brotzge, J., & Donner, W. (2013). The tornado warning process: A review of current research, challenges, and opportunities. *Bulletin of the American Meteorological Society*, *94* (11), 1715-1733.

Budescu, D. V., Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic change*, *113* (2), 181-200.

Burnside, R. (2006). Leaving the big easy: An examination of the hurricane evacuation behavior of New Orleans residents before Hurricane Katrina. *Journal of Public Management and Social Policy*, *12* (2), 49-61.

Carbone, G. J., & Dow, K. (2005). Water resource management and drought forecasts in South Carolina. *JAWRA Journal of the American Water Resources Association*, *41* (1), 145-155.

Center for Research on Environmental Decisions (CRED) (2009). *The Psychology of Climate Change Communication: A Guide for Scientists, Journalists, Educators, Political Aides, and the Interested Public*. New York.

Chabris, C. F., & Simons, D. J. (2011). *The Invisible Gorilla: How Our Intuitions Deceive Us*. Random House LLC.

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P., & Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8* (2), 024024.

Cotté, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger drivers' reliance on collision warning systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45* (4), 277-280. SAGE Publications.

Cuite, C. L., Weinstein, N. D., Emmons, K., & Colditz, G. (2008). A test of numeric formats for communicating risk probabilities. *Medical Decision Making, 28* (3), 377-384.

Cutter, S. L., & Smith, M. M. (2009). Fleeing from the hurricane's wrath: Evacuation and the two Americas. *Environment: Science and Policy for Sustainable Development, 51* (2), 26-36.

Dash, N., & Gladwin, H. (2007). Evacuation decision making and behavioral responses: Individual and household. *Natural Hazards Review, 8* (3), 69-77.

Diehl, E., & Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes, 62* (2), 198-215.

Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., & Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions of scientific agreement. *Nature Climate Change, 1*, 462-466.

Dow, K., & Cutter, S. L. (1998). Crying wolf: Repeat responses to hurricane evacuation orders. *Coastal Management, 26* (4), 237-252.

Dow, K., & Cutter, S. L. (2000). Public orders and personal opinions: Household strategies for hurricane risk assessment. *Environmental Hazards, 2,* 143-155.

Earle, T. C., & Siegrist, M. (2006). Morality information, performance information, and the distinction between trust and confidence. *Journal of Applied Social Psychology, 36* (2), 383-416.

Eckel, F. A., Allen, M. S., & Sittel, M. C. (2012). Estimation of ambiguity in ensemble forecasts. *Weather & Forecasting, 27* (1), 50-69.

Edwards, A., Elwyn, G., Covey, J., Matthews, E., & Pill, R. (2001). Presenting risk information a review of the effects of framing and other manipulations on patient outcomes. *Journal of health communication, 6* (1), 61-82.

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment*, 17-52.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual review of psychology*, *32* (1), 53-88.

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal*, *324* (7339), 729-732.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112* (4), 912-931.

Erickson, S. A., & Brooks, H. (2006). Lead time and time under tornado warnings: 1986–2004. In *23rd Conference on severe local storms.* St. Louis, MO.

Ert, E., & Erev, I. (2008). The rejection of attractive gambles, loss aversion, and the lemon avoidance heuristic. *Journal of Economic Psychology*, *29* (5), 715-723.

Finucane, M.L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13,* 1-17.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely=rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making*, *9* (3), 153-172.

Fischhoff, B. (1995). Risk perception and communication unplugged: Twenty years of process. *Risk Analysis, 15* (2), 137-145.

Fischhoff, B., Bostrom, A., & Quadrel, M. J. (1993). Risk perception and communication. *Annual Review of Public Health, 14,* 183-203.

Forer, B., Tanglao, L., Schabner, D. (2011). Fall storm: October Nor'Easter blamed for at least three deaths. Accessed online on March 5, 2012, at http://abcnews.go.com/US/fall-storm-october-noreaster-leaves-dead/story?id=14838651.

Freudenburg, W. R., & Muselli, V. (2010). Global warming estimates, media expectations, and the asymmetry of scientific challenge. *Global Environmental Change, 20,* 483-491.

Frewer, L. J., Hunt, S., Brennan, M., Kuznesof, S., Ness, M., & Ritson, C. (2003). The views of scientific experts on how the public conceptualize uncertainty. *Journal of Risk Research, 6* (1), 75-85.

Gallup (2014). Party affiliation. Accessed online on February 1, 2014, at http://www.gallup.com/poll/15370/party-affiliation.aspx.

Garcia-Retamero, R., & Galesic, M. (2012). Doc, what would you do if you were me? On self–other discrepancies in medical decision making. *Journal of Experimental Psychology: Applied, 18* (1), 38.

Gärling, T., Kirchler, E., Lewis, A., & van Raaij, F. (2009). Psychology, financial decision making, and financial crises. *Psychological Science in the Public Interest, 10* (1), 1-47.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a

warning on human response time. *Journal of Experimental Psychology: Applied*, *1* (1), 19.

Gibbs, L. (2002). *Aesop's Fables.* A new translation by Laura Gibbs. Oxford University Press (World's Classics): Oxford. Accessed online on February 1, 2014 at http://mythfolklore.net/aesopica/oxford/151.htm.

Gibbs, L. I., & Holloway, C. F. (2013). *Hurricane Sandy After Action: Report and Recommendations to Mayor Michael R. Bloomberg*. Accessed online on February 1, 2014 at http://www.nyc.gov/html/recovery/downloads/pdf/sandy_aar_5.2.13.pdf.

Gigerenzer, G. (2008). *Gut feelings: Short cuts to better decision making*. Penguin UK.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, *327*, 741-744.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological review*, *102* (4), 684-704.

Gilovich, T., & Griffin, D. (2002). Introduction-heuristics and biases: Then and now. *Heuristics and biases: The psychology of intuitive judgment*, 1-18.

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310* (5746), 248-249.

Gruntfest, E., Downing, T., and White, G. F. (1978). Big Thompson Flood. Working Paper No. 32, Institute of Behavioral Science, Univ. of Colorado, Boulder, CO.

Gupta, N., Bisantz, A. M., & Singh, T. (2001). Investigation of factors affecting driver performance using adverse condition warning systems. In *Proceedings of the Human*

*Factors and Ergonomics Society Annual Meeting, 45* (23), 1699-1703. SAGE Publications.

Hallinan, J. T. (2010). *Why we make mistakes: How we look without seeing, forget things in seconds, and are all pretty sure we are way above average*. Random House LLC.

Hamrick, D. (2011). Storm summary number 3 for autumn Mid-Atlantic to northeast U.S. major winter storm (NWS Hydrometeorological Prediction Center). Accessed online on March 5, 2012, at http://www.webcitation.org/62mgX5qcg.

Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition, 110*, 51-64.

Hart, D., & Polson, J. (2011). More than 2M without power after U.S. snow. Accessed online on March 5, 2012, at http://www.bloomberg.com/news/2011-10-31/about-3-million-without-power-as-freeze-to-hit-u-s-northeast.html.

Hastie, R., & Dawes, R. M. (eds.). (2010). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage.

Hau, R., Pleskac, T. J., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making, 23* (1), 48-68.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15* (8), 534-539.

Hoefer, C. (2008). Causal determinism. Accessed online on February 1, 2014, at http://plato.stanford.edu/archives/win2009/entries/determinism-causal/.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73 (3)*, 538-540.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84* (3), 343-352.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *The American Economic Review, 95* (3), 902-904.

Jacoby, J., Mazursky, D., Troutman, T., & Kuss, A. (1984). When feedback is ignored: Disutility of outcome feedback. *Journal of Applied Psychology*, *69* (3), 531-545.

Johnson, B. B., & Slovic, P. (1995). Presenting uncertainty in health risk assessment: Initial studies of its effects on risk perception and trust. *Risk Analysis*, *15*, 485-494.

Johnson, B. B., & Slovic, P. (1998). Lay views on uncertainty in environmental health risk assessment. *Journal of Risk Research*, *1*, 261-279.

Joslyn, S., Nadav-Greenberg, L. & Nichols, R.M. (2009). The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Applied Cognitive Psychology, 23*, 55-72.

Joslyn, S. & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, *17*, 180-195.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *9*, 697-720.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64 (6),* 515-526.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47* (2), 263-292.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39* (4), 341-350.

Kerstholt, J. H. (1995). Decision making in a dynamic situation: The effect of false alarms and time pressure. *Journal of Behavioral Decision Making, 8* (3), 181-200.

Kerstholt, J. H., & Passenier, P. O. (2000). Fault management in supervisory control: The effect of false alarms and support. *Ergonomics, 43*(9), 1371-1389.

Klein, G. (2009). *Streetlights and Shadows*. Boston: MIT Press.

Klein, G. (2013). *Seeing What Others Don't: The Remarkable Ways We Gain Insights*. New York: Public Affairs.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin, 119* (2), 254-284.

Kootval, H. (2008). Guidelines on Communicating Forecast Uncertainty. *World Meteorological Organization/Technical Document* (4122).

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual review of psychology, 50* (1), 569-598.

Leiserowitz, A., Maibach, E., Roser-Renouf, C., Feinberg, G., & Howe, P. (2013). *Climate change in the American mind: Americans' global warming beliefs and attitudes in April*

*2013*. Yale University and George Mason University. New Haven, CT: Yale Project on Climate Change Communication.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making, 27* (5), 696-713.

Luo, L., & Pan, M. (2013). Drought monitoring and hydrologic forecasting with VIC. Accessed online on February 2, 2014, at http://hydrology.princeton.edu/forecast/current.php.

Malenka, D. J., Baron, J. A., Johansen, S., Wahrenberger, J. W., & Ross, J. M. (1993). The framing effect of relative and absolute risk. *Journal of General Internal Medicine, 8* (10), 543-548.

McKinley, J. C., & Urbina, I. (2008). A million flee as huge storm hits Texas coast. *The New York Times*. Accessed online on September 10, 2010, at http://query.nytimes.com/gst/fullpage.html?res=9D05EFDF143EF930A2575AC0A96E9C8B63.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota Press.

Miles, S., & Frewer, L. J. (2003). Public perception of scientific uncertainty in relation to food hazards. *Journal of risk research, 6* (3), 267-283.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Weather and Forecasting, 23*, 974-991.

Morss, R. E., Lazo, J. K., & Demuth, J. L. (2010). Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication. *Meteorological Applications, 17* (2), 149-162.

Moser, S. C. (2010). Communicating climate change: History, challenges, process and future directions. *Climate Change, 1*, 31-53.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5* (1), 2-12.

Murphy, A. H. (1991). Probabilities, odds, and forecasts of rare events. *Weather and forecasting*, *6* (2), 302-307.

Murphy, A. H. (1998). The early history of probability forecasts: Some extensions and clarifications. *Weather & Forecasting*, *13* (1), 5-15.

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society*, *61* (7), 695-701.

Murphy, A. H., & Winkler, R. L. (1979). Probabilistic temperature forecasts: The case for an operational program. *Bulletin of the American Meteorological Society, 60* (1), 12-19.

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among non-experts. *Journal of Cognitive Engineering and Decision Making, 3* (3), 209-227.

Nagele, D. E., & Trainor, J. E. (2012). Geographic specificity, tornadoes, and protective action. *Weather, Climate & Society*, *4*(2), 145-155.

National Research Council (NRC). (2006). *Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts.* Washington, DC: National Academies Press.

NOAA (2011). NWS Central Region Service Assessment: Joplin, Missouri, Tornado – May 22, 2011. Accessed online on October 26, 2013, at http://www.nws.noaa.gov/om/assessments/pdfs/Joplin_tornado.pdf.

NOAA (2012a). National Weather Service Drought Fact Sheet. Accessed online on August 10, 2013, at http://www.nws.noaa.gov/om/csd/graphics/content/outreach/brochures/FactSheet_Drought.pdf.

NOAA (2012b). National Climatic Data Center: Drought, June 2012. Accessed online on August 10, 2013, at http://www.ncdc.noaa.gov/sotc/drought/2012/6.

NOAA (2013). Service Assessment: Hurricane/Post-Tropical Cyclone Sandy, October 22-29, 2012. Accessed online on October 26, 2013, at http://www.nws.noaa.gov/os/assessments/pdfs/Sandy13.pdf.

Onishi, N., & Wollan, M. (2014). Severe drought grows worse in California. Accessed online on February 2, 2014, at http://www.nytimes.com/2014/01/18/us/as-californias-drought-deepens-a-sense-of-dread-grows.html?_r=0.

Palmer, T. N., & Räisänen, J. (2002). Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature*, *415* (6871), 512-514.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, *5* (5), 411-419.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*( 2), 230-253.

Parker, D. J., Priest, S. J., & Tapsell, S. M. (2009). Understanding and enhancing the public's behavioural response to flood warning information. *Meteorological applications, 16* (1), 103-114.

Patt, A. (2001). Understanding uncertainty: Forecasting seasonal climate for farmers in Zimbabwe. *Risk Decision and Policy, 6* (2), 105-119.

Patt, A. G., & Schrag, D. P. (2003). Using specific language to describe risk and probability. *Climatic Change, 61*, 17-30.

Paul, B. K., & Stimers, M. (2012). Exploring probable reasons for record fatalities: The case of 2011 Joplin, Missouri, Tornado. *Natural hazards, 64* (2), 1511-1526.

Payne, J. W. (2005). It is whether you win or lose: The importance of the overall probabilities of winning or losing in risky choice. *Journal of Risk and Uncertainty, 30* (1), 5-19.

Perry, R. W. (1983). Population evacuation in volcanic eruptions, floods, and nuclear power plant accidents: Some elementary comparisons. *Journal of Community Psychology, 11* (1), 36-47.

Polman, E. (2010). Information distortion in self-other decision making. *Journal of Experimental Social Psychology, 46* (2), 432-435.

Riad, J. K., Norris, F. H., & Ruback, R. B. (1999). Predicting evacuation in two major disasters: Risk perception, social influence, and access to resources. *Journal of Applied Social Psychology, 29* (5), 918-934.

Ritov, I., & Baron, J. (1995). Outcome knowledge, regret, and omission bias. *Organizational Behavior and Human Decision Processes, 64* (2), 119-127.

Roulston, M. S., Bolton, G. E., Kleit, A. N., & Sears-Collins, A. L. (2006). A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting, 21*, 116-122.

Roulston, M. S., & Smith, L. A. (2004). The boy who cried wolf revisited: The impact of false alarm intolerance on cost–loss scenarios. *Weather & Forecasting*, *19* (2), 391-397.

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38* (7), 608-625.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, *53* (2), 252-266.

Sheffield, J., Wood, E. F., & Roderick, M. L. (2012). Little change in global drought over the past 60 years. *Nature*, *491* (7424), 435-438.

Shepherd, M. (2014). An open thank you to meteorologists in Atlanta. Accessed online on February 16, 2014, at http://meteorologistandtheatlantasnow2014.blogspot.com/2014/01/an-open-thank-you-letter-to-atlanta.html.

Silver, N. (2012a). *The signal and the noise: Why so many predictions fail-but some don't*. New York: The Penguin Press.

Silver, N. (2012b). The weatherman is not a moron. Accessed online on April 21, 2014, at http://www.nytimes.com/2012/09/09/magazine/the-weatherman-is-not-a-moron.html?pagewanted=all&_r=0.

Simon, H. A. (1957). *Models of man; social and rational*. Oxford, England: Wiley.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin,* 119 (1), 3-22.

Sloughter, J. M., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, *135* (9), 3209-3220.

Slovic, P. (1987). Perception of risk. *Science, New Series, 236* (4799), 280-285.

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk analysis*, *13* (6), 675-682.

Slovic, P. (1999). Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. *Risk Analysis, 19 (4)*, 689-701.

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* (2), 311-322.

Slovic, P., Monahan, J., & MacGregor, D. M. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instructions, and employing probability vs. frequency formats. *Law and Human Behavior, 24* (3), 271-296.

Sorensen, J. H. (2000). Hazard warning systems: Review of 20 years of progress. *Natural Hazards Review*, 119-125.

Stanovich, K. E., & West, R. F. (2002). Individual differences in reasoning: Implications for the rationality debate. In T. Gilovich, D. Griffin & D. Kahneman (eds.), Heuristics and biases (pp. 421–440). Cambridge: Cambridge University Press.

Stone, E. R., Yates, J. F., & Parker, A. M. (1994): Risk communication: Absolute versus relative expressions of low-probability risks. *Organizational Behavior and Human Decision Processes, 60,* 387-408.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thompson, J. C. (1952). On the operational deficiencies in categorical weather forecasts. *Bulletin of the American Meteorological Society, 33,* 223-226.

Tindale, R. S. (1989). Group versus individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes, 44* (3), 454-473.

Trenberth, K. E. (2012). Framing the way to relate climate extremes to climate change. *Climatic Change, 115* (2), 283-290.

Trepel, C., Fox, C. R., & Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive Brain Research, 23* (1), 34-50.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185* (415), 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211* (4481), 453-458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty, 5* (4), 297-323.

U.S. Census Bureau (2014). Selected population profile in the United States: 2012 American community survey 1-year estimates. Accessed online on February 1, 2014, at

http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_1YR_S0201&prodType=table.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115* (4), 348-365.

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, *77* (1-2), 103-120.

Weber, R. A. (2003). 'Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior*, *44* (1), 134-144.

Whitehead, J. C., Edwards, B., Van Willigen, M., Maiolo, J. R., Wilson, K., & Smith, K. T. (2000). Heading for higher ground: Factors affecting real and hypothetical hurricane evacuation behavior. *Global Environmental Change Part B: Environmental Hazards*, *2* (4), 133-142.

Windschitl, P. D., & Weber, E. (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 & Cognition*, 25 (6), 1514-1533.

Wu, G., Zhang, J., & Gonzalez, R. (2004). Decision under risk. In *Blackwell Handbook of Judgment and Decision Making* (Koehler, D. J., & Harvey, N., eds.), 399-424.

Yamagishi, K. (2007). When a 12.86% mortality is more dangerous than 24.14%: Implications for risk communication. *Applied Cognitive Psychology*, *11*, 495-506.

Zhu, Y., & Toth, Z. (2001). Extreme weather events and their probabilistic prediction by the NCEP Ensemble Forecast System. Preprints, Symposium on Precipitation Extremes: Prediction, Impact, and Responses. Albuquerque, NM, American Meteorological Society, 82-85.

Ziegler, F. V., & Tunney, R. J. (2012). Decisions for others become less impulsive the further away they are on the family tree. *PloS one*, *7* (11), e49479.