Journal of Experimental Psychology: Applied
2012, Vol. 18, No. 1, 126–140

1076-898X/11/$12.00 DOI: 10.1037/a0025185

Uncertainty Forecasts Improve Weather-Related Decisions and Attenuate the Effects of Forecast Error

Susan L. Joslyn and Jared E. LeClerc
University of Washington

Although uncertainty is inherent in weather forecasts, explicit numeric uncertainty estimates are rarely included in public forecasts for fear that they will be misunderstood. Of particular concern are situations in which precautionary action is required at low probabilities, often the case with severe events. At present, a categorical weather warning system is used. The work reported here tested the relative benefits of several forecast formats, comparing decisions made with and without uncertainty forecasts. In three experiments, participants assumed the role of a manager of a road maintenance company in charge of deciding whether to pay to salt the roads and avoid a potential penalty associated with icy conditions. Participants used overnight low temperature forecasts accompanied in some conditions by uncertainty estimates and in others by decision advice comparable to categorical warnings. Results suggested that uncertainty information improved decision quality overall and increased trust in the forecast. Participants with uncertainty forecasts took appropriate precautionary action and withheld unnecessary action more often than did participants using deterministic forecasts. When error in the forecast increased, participants with conventional forecasts were reluctant to act. However, this effect was attenuated by uncertainty forecasts. Providing categorical decision advice alone did not improve decisions. However, combining decision advice with uncertainty estimates resulted in the best performance overall. The results reported here have important implications for the development of forecast formats to increase compliance with severe weather warnings as well as other domains in which one must act in the face of uncertainty.

Keywords: risk-seeking, decision making, uncertainty

Many important decisions are made under uncertainty, such as choosing medical treatment, choosing retirement investments, and deciding whether to take precautionary action against severe weather. It is now possible in several domains to quantify uncertainty, providing accurate numeric assessments to end-users. This is especially true of weather forecasts. Recent technical advancements have resulted in methods for calculating forecast probabilities that are both calibrated and precise (Gneiting & Raftery, 2005; Sloughter, Raftery, Gneiting, & Fraley, 2007). However, very little of this information reaches the general public. At present, most public weather forecasts, like information provided to nonexperts in a broad range of domains, remain deterministic. They provide a single value, such as a nighttime low temperature of 32 °F, implying unrealistic certainty.

Although there are strong theoretical arguments for the economic benefit of uncertainty forecasts (Murphy, 1977; Thompson, 1952) and although psychological research suggests that numeric estimates lead to a more precise understanding than do verbally described categories (Budescu, Broomel & Por, 2009), there is still widespread reluctance to provide numeric uncertainty to the general public for fear it will be misused. Indeed, many of the crucial decisions based on weather forecasts are similar to those that have been shown in laboratory studies to lead to a decision error referred to as risk-seeking (Kahneman & Tversky, 1979). Weather-related decisions often concern costly precautionary action and must be made early, when the probability of adverse weather is well below 50%. Experimental participants asked to choose between options that involve losses tend to prefer a gamble with a moderately low probability, even when the probability-weighted value, called the expected value (Bernoulli, 1954), is less than the certain alternative (Kahneman & Tversky, 1984). In laboratory settings in which all of the relevant information is provided to the participant, this error, referred to as “risk seeking,” and the opposite “risk averse” error, preferring a sure option to a risky option with equal or greater expected value, can be explained by cumulative prospect theory (Tversky & Kahneman, 1992). According to cumulative prospect theory, both errors result from a combination of the effects of a nonlinear utility function, which is steeper for losses than for gains, and a nonlinear subjective probability weighting function. We focus here on the risk-seeking error, which has been demonstrated in diverse domains including medicine (Nightingale, 1987), used car purchasing (Betts & Taran, 2005), and international relations (Haas, 2001). Such evidence suggests that, in a variety of situations involving loss, people assume more risk than is economically rational. This is true of both professionals as well as nonexperts, suggesting that it is not merely a matter of lack of education or experience.

This article was published Online First August 29, 2011.

Susan L. Joslyn and Jared E. LeClerc, Department of Psychology, University of Washington.

This research was supported by the National Science Foundation, Under Grant ATM 0724721.

Correspondence concerning this article should be addressed to Susan L. Joslyn, Department of Psychology, Box 351525, University of Washington, Seattle, WA 98195. E-mail: susanj@u.washington.edu

It is important to note, however, that these conclusions are based on comparisons between human decision makers and models of economically rational choice. If the comparison is made instead between human decisions made with and without explicit uncertainty estimates, the conclusion is somewhat different. People make *better* decisions with explicit uncertainty estimates than without them in a number of complex naturalistic decision tasks when they are given sufficient practice and motivation (Joslyn, Pak, Jones, Pyles, & Hunt, 2007; Nadav-Greenberg & Joslyn, 2009; Roulston, Bolton, Kleit, & Sears-Collins, 2006).

In one study (Nadav-Greenberg & Joslyn, 2009), participants used weather forecasts to make decisions about treating the roads to prevent icing. Some participants received a deterministic forecast, the overnight low temperature, while others also received the probability of freezing. Both groups chose to treat the roads more often when the probability of freezing was high and less often when it was low, suggesting that even when it was not explicit, participants estimated the likelihood of freezing temperatures. In fact, the nighttime low temperature was highly correlated with the probability of freezing, as real forecast data was used.

In this task, the cost of treatment was $1,000 and the penalty for nontreatment when freezing temperatures occurred was $6,000, so the optimal strategy was to salt whenever the probability of freezing exceeded 17% (1000/6000) at which point the expected value of the penalty exceeded the cost of treatment. By this standard, both groups tended to be risk-averse, salting too often below the optimal threshold, *and* risk-seeking, salting too little above the threshold. However, those with probability forecasts made fewer errors of both kinds. Furthermore, it was clear that participants realized the advantage of the explicit uncertainty estimate. In a condition in which participants were required to request the type of forecast information they wanted to use, all of them chose some form of uncertainty information and did so on 85% of the trials, suggesting that they understood the value of such information.

Although the participants using probabilistic forecasts in this experiment were functionally less risk-seeking and less risk-averse, their advantage over those using deterministic forecasts may have been due to the fact that they had an accurate assessment of the risk. There is now good evidence that people attempt to estimate forecast uncertainty when it is not provided, in some cases making serious errors (Joslyn & Savelli, 2010). For instance, people think that extreme single-value forecasts, such as a daytime high in Seattle of 100 °F, are likely to verify much closer to normal values, when in fact the opposite is true. Thus, in the experiment described above, the well-calibrated probabilistic forecasts may simply have provided participants with a more accurate risk assessment than they could estimate on their own, allowing them to make better decisions.

Nonetheless, there may be some weather events for which quantifying the uncertainty is *not* advantageous. For instance, severe weather warnings must be issued sufficiently early to allow for successful evacuation. Often, a few days in advance, the probability of adverse weather in a particular geographic region is less than 20%. Emergency managers worry that specifying the probability in such cases may make the event seem *less* likely to people and decrease rather than increase compliance with warnings. To be sure, in the study described above (Nadav-Greenberg & Joslyn, 2009), despite the overall improvement with probabilistic forecasts, the advantage did not extend to the lowest probabilities of freezing for which applying treatment was economically rational (17%−20%). However, this study did not test a condition in which people were given explicit advice about what to do, as they are with weather warnings. Hence it is unclear how decisions based on uncertainty forecasts would compare with being told what to do, without being told why. As far as we know, no research tests this critical question.

Indeed, compliance with weather warnings under the current system is often poor. For example, in September 2008 the residents of Galveston, Texas, were warned that they faced "certain death" as a result of Hurricane Ike if they failed to comply with mandatory evacuation orders, and yet noncompliance was estimated at 40% (McKinley & Urbina, 2008). This troubling noncompliance rate may be due to the risk-seeking tendency described above. Residents often regard evacuation as costly (Blendon, 2008; Cutter & Smith, 2009), fitting the profile of choices that elicit such errors.

However, noncompliance with weather warnings may be due instead to the fact that residents do not have a clear understanding of the situation. Without reliable uncertainty estimates, they may underestimate the likelihood of severe weather. In addition, they may not realize the severity of the impact of the adverse weather on their lives, mistakenly thinking that the cost of evacuation is greater. In other words, they may not be sufficiently worried about it. Similar lack of affect is suspected in failure to pursue risk management in climate change issues (Weber, 2006). For weather warnings, such perceptions may be enhanced by distrust in the forecast due to prior experience with false alarms. Because of the low probability at which preparation is required, warning forecasts tend to result in a high false alarm rate (Joslyn & Savelli, 2010; Barnes, Gruntfest, Hayden, Schultz, & Benight, 2007). Obviously, from the forecasters' perspective, a miss is the preferred error because it is usually less costly than a false alarm, at least in terms of the immediate outcome. However, this reasoning is not made clear to the end-user.

Indeed, the long-term psychological effects of forecast error are unclear. There is some evidence suggesting that people are fairly tolerant of error. In one study, for instance, the majority of hurricane evacuees who experienced a false alarm expressed willingness to evacuate in the future (Baker, 1991). However, other evidence from a number of domains suggests that false alarms lead to distrust (Gupta, Bisantz, & Singh, 2001; Kahn & Luce, 2003), ignoring future alarms (Lees & Lee, 2007; Parasuraman & Riley, 1997; Nohre, 2001), and delayed response to alarms (Kerstholt, 1995). Furthermore, some claim that once lost, trust is difficult to regain (Slovic, 1999). If so, potential negative effects of forecast error may be long lasting.

In sum, although there is a distinct advantage for uncertainty forecasts in a number of situations when compared to decisions based on deterministic forecasts, whether or not this advantage extends to situations in which action is required at low probabilities or when forecast error increases is as yet unknown. We also do not yet know whether uncertainty forecasts lead to better decisions than giving people explicit advice about what to do. The purpose of the research reported here is to address these questions. In doing so we will also explore the cognitive mechanisms that give rise to risk-seeking errors in complex naturalistic situations and recommend methods for communicating risk to overcome this decision error.

These experiments employ the basic road salt paradigm described above. This experimental paradigm has several advantages in terms of ecological validity. Participants are informed of the observed temperature after each of 120 forecasts, gaining experience with the relationship between the forecast and its outcome. In addition, the decisions have actual consequences. Participants are paid cash commensurate with their performance. Although the cover story for the experimental task involves a job that would in fact be done by a professional, it is used here to provide meaningful motivation to a repeated decision task, simulating the typical binary decision about taking precautionary action required for most weather events. The participants themselves were nonexperts. Thus, even though there are clearly many differences between this task and public response to weather warnings, the fact that they have similarly low rational probability thresholds for action, that precautionary action involves costs in both cases, and that the decision makers are nonexperts allows us to use this paradigm to explore some of the causes for risk-seeking errors in natural settings. Finally, this paradigm involves clear-cut “correct” responses from an economic perspective, allowing for firm conclusions about improvement in performance.

This experimental paradigm will allow us to test two main questions addressing the psychological causes for decision errors in a realistic context. First, are risk-seeking errors due to misunderstanding the situation? If so, then uncertainty expressions emphasizing the potential loss may spur people to action even when the probability is low. Two manipulations addressed this hypothesis. In Experiment 1, frequency expressions were tested, previously shown to increase the salience of rare events (Koehler, 2001; Slovic, Mohnahan & McGregor, 2000). In Experiment 2, the potential penalty was emphasized on every forecast. Although forecasts emphasizing end-user consequences (e.g., widespread power outages) are often recommended (Seigrist & Gutscher, 2008), as far as we are aware, this suggestion has never been tested experimentally. A decrease in risk-seeking errors with either of these manipulations suggests that such errors reside in the comparison of the two potential outcomes.

The other main question is whether risk-seeking errors arise from distrust in the forecast due to prior experience with forecasts perceived to be inaccurate. To answer this question, in Experiments 1 and 3, forecast error was systematically manipulated to determine whether greater error reduced trust and willingness to take precautionary action. Manipulating forecast error had the added benefit of allowing us to test whether forecasts that include uncertainty estimates provide protection against any negative effects of forecast error, as is often claimed (National Research Council, 2006). The notion is that a forecast acknowledging the uncertainty will seem less “wrong” when the single-value forecast fails to verify, preserving user trust. However, as far as we know, this claim has never been tested systematically.

To provide a thorough test of these hypotheses, in all three experiments, we also included conditions that are directly comparable to the weather-warning situation in which people are informed of the appropriate course of action (e.g., mandatory evacuation). In these conditions participants were given explicit recommendations about the economically rational course of action based on the expected loss for that trial. In Experiment 1, the advice was provided without explanation, and in Experiment 2, its derivation was also explained in an attempt to enhance compliance. This allowed us to compare conditions in which people were given good advice without an uncertainty estimate (as they are in evacuation notices) to conditions in which people were provided with the relevant uncertainty information and allowed to make their own decisions. Experiment 3 combined the advice with an uncertainty estimate to test whether benefits observed for each format could be realized in a single forecast expression.

Experiment 1

Experiment 1 tested four forecast formats to determine the relative advantage of augmenting the conventional deterministic forecast with either uncertainty or explicit advice when forecast error increased. We hypothesized a general decline in performance in high-error trials. We expected those with explicit advice to perform better than those using only the deterministic forecast. We expected those with uncertainty forecasts to perform better overall, express greater trust in the forecast and experience less decline due to forecast error. We also tested uncertainty expressed as frequency, expecting an advantage for this format over the probabilistic expression in the low probability of freezing trials.

Method

Participants. A total of 304 University of Washington psychology students (50% female) participated for course credit and the chance to earn prize money. Mean age was 19.5 years (range 16–50).

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

Procedure. After participants gave informed consent and entered demographic information, they read a set of instructions at the same time that the experimenter read them aloud, which included a description of the task and the cost-loss structure. Participants were to assume the role of a president of a road maintenance company contracted to treat the roads in a U.S. town with salt brine to prevent icing. There were 120 trials representing four late fall and winter months. Participants received a virtual monthly budget of $36,000. Applying salt brine cost $1,000 per day. The penalty for failing to apply salt brine when a freezing temperature was observed was $6,000. Participants were instructed to attempt to maximize profits by minimizing salting expenses and avoiding penalties. At the end of the experiment, participants were paid $1 for every $1,000 in their final balance over $24,000, such that applying salt brine on every trial would constitute breaking even.

In each trial, representing one day, a forecast for the next night appeared on the screen. Participants rated their trust in the forecast on a 5-point drop-down menu, ranging from “very little” to “very much.” Next, they clicked on one of two boxes marked “Salt” or “Not salt.” Finally, participants entered a numeric value in a text box to indicate what they thought the nighttime low temperature would be. Immediately afterward, the observed nighttime low temperature and any budget adjustments appeared on the screen. Participants were able to borrow against the next month’s budget installment if their balance dropped below $0. After 30 trials, representing one month, participants again indicated their overall trust in the forecast. Participants clicked “Next” to continue on to

the next month's trials, and $36,000 was added to the budget. At the end of the final "month," participants with ending budgets above $24,000 were paid and all participants were awarded course credit points. Experimental sessions included one to 12 participants and lasted approximately 45 minutes.

Design. A 4 (between) $\times$ 2 (within) mixed-model design was used. The between-groups variable was forecast format. Participants were randomly assigned to one of four forecast formats, all of which included the deterministic forecast. The control condition included only the deterministic forecast. In the freeze probability condition uncertainty was expressed as percent chance (e.g., "there is a 22% chance that the temperature will be equal to or less than 32 degrees"). In the Freeze frequency condition uncertainty was expressed as the number of days out of 100 (e.g., "22 out of 100 days like this the temperature will be equal to or less than 32 degrees"). In the "decision aid" condition, participants were informed that an advanced computer modeling system, combining information about the forecast, the uncertainty, and the cost of salting and penalty for not salting, would provide salting recommendations (e.g., "Applying salt brine is recommended under these circumstances."). In fact, it recommended applying salt brine whenever the probability of freezing was equal to or greater than the economically rational threshold 17% (1000/6000).

The within-groups variable was the difference between the temperature forecast and the observation, the amount of error. One block of 60 trials had a mean standard error of 3.26 °F (low-error trials) and the other had a mean standard error of 6.53 °F (high-error trials). In both blocks the error was unbiased. Blocks were counterbalanced such that half of participants had the low-error forecasts first; the other half had the high-error forecasts first. The dependent variables were final balance, a measure of decision rationality referred to as expected loss explained below, trust rating and binary salt decision.

Stimuli. Participants in all four conditions received the same deterministic night time low temperature forecasts and observed temperatures in the same order within blocks. The ranges of temperature, probabilities of freezing, and forecast error in these 60 forecasts were based on historical forecast data from the cities of Spokane and Yakima in Washington State. The deterministic forecasts ranged from 32 °F to 37 °F ($M = 34$ °F). The probabilities of freezing (PoF) ranged from 10% to 51% ($M = 24.75\%$) and were drawn from normal distributions with means representing the deterministic forecasts. Observed temperatures ranging from 27 °F to 42 °F ($M = 35$ °F) were generated using Excel's random number function to select values between 0 and 1, representing the portion beneath the curve of each distribution. Then, using historic weather data as a guide a typical sequence of weather events was created, with observed temperatures following trends and including only natural fluctuations from one night to the next (<16 °F).

The probabilistic forecasts were well calibrated. Half of all observed temperatures were above their respective deterministic temperature forecasts and half were below. Probability forecasts were divided into six range categories (10–16%, 17–23%, 24–33%, 31–37%, 38–44%, 45–51%). The percentage of observed temperatures 32 °F or less in each category was within that probability range. For example, in the 10–16% range, temperatures 32 °F or less were observed on 2 of 18 (11.1%) days. There were 18 forecasts in both the 10–16% and the 17–23% ranges to allow for decision comparisons in these two categories of particular interest.

Forecast error was manipulated in blocks of trials. The low-error block comprised 60 forecasts with a mean standard error of 3.26 °F. The high-error forecasts were created from the low-error forecasts by holding the deterministic temperature constant and doubling the standard deviations of each distribution to provide a new set of observations (ranging from 21 °F to 51 °F; $M = 35$ °F) resulting in greater error ($SE = 6.53$ °F). In addition, although the deterministic forecasts in the high-error trials were identical to those in the low-error trials, the increased standard deviations in the distributions changed the probabilities of freezing, resulting in a range from 26% to 51% ($M = 36\%$).

Results

Although our main hypotheses concerned decision quality, we first examined participants' numeric temperature estimates to determine whether they understood the stimuli. In order to summarize across the differing forecasts, we subtracted the deterministic forecast from participants' numeric estimate for each trial and calculated means for the low- and high-error trials separately. There were 20 participants with values at least two standard deviations above the mean standard error in their condition who were omitted from the analyses below.

Among the remaining 284 participants, the order in which the forecast error blocks were presented impacted many dependent variables. Thus, orders are analyzed separately below to highlight these differences. In addition, subsequent analyses omit trials in the last "month" of the simulation because, on average, participants finished with balances below the minimum necessary to earn a cash prize; indeed, this was the case with all forecast formats except freeze frequency. Participants with low balances might have salted less in the final (4th) month of the simulation, when no further budget installments could be anticipated, for that reason. To eliminate this uncontrolled variable we omitted both Months 2 and 4 for the analyses below in which we compare low- and high-error trials in order to compare months with identical forecasts, Months 1 and 3. All effect sizes are reported using Cohen's *d*.

Expected loss. Our expectations for higher ending balances among those with uncertainty information and in low-error forecasts were confirmed (see Table 1). However, it must be noted that balance is not a pure measure of decision quality because it was affected in part by chance. For instance, a participant might make an economically rational decision, choosing not to salt when the PoF was less than 17%, but be penalized $6,000 because of a rare observed temperature below freezing. A direct measure of the economic value of decision-making is expected loss (EL) based on the idea of expected value (Bernoulli, 1954). Expected loss was the penalty (−$6,000) multiplied by the PoF for each trial on which participants decided not to salt. Participants who decided to salt were assigned the cost of salting (−$1,000). A mean EL was calculated for the first months of the low- and high-error blocks for each participant. To account for the fact that economically optimal decision-making led to a different mean EL for the low- (−$924) and high-error (−$1,000) trials, the appropriate optimal mean EL was subtracted from the participants' mean EL for each month (see Table 2). A mixed-model ANOVA was conducted on expected loss difference with forecast format (freeze frequency, freeze prob-

Table 1
Experiment 1: Mean Balance Difference Scores (Participant Balance–Optimal Balance) on Low-Error-First Trials

Forecast Format	Low-error trials *M* (*SD*)	High-error trials *M* (*SD*	*Overall*
Control	−\$2257.14 (\$5967.28)	−\$13057.14 (\$9579.21)	*−\$7657.14*
Decision Aid	−\$1461.54 (\$5812.18)	−\$10358.97 (\$9702.11)	*−\$5910.26*
Freeze Frequency	\$1189.19 (\$3942.91)	−\$5324.32 (\$7287.94)	*−\$2067.57*
Freeze Probability	\$3102.56 (\$4147.26)	−\$7128.21 (\$8121.38)	*−\$2012.82*
Overall	*\$193.33* (*\$5428.85*)	*−\$8906.67* (*\$9122.10*)	*−\$4411.95*

Note. Higher difference scores suggest better performance.

ability, decision aid, and control) as the between-groups independent variable and forecast error (low and high) as the within-groups independent variable. For participants who encountered the low forecast error block first, there was 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 ($M = -\$204.70$, $SD = \$142.17$) than either control participants ($M = -\$330.47$, $SD = \$141.35$, $p < .01$, $d = .89$) or decision aid participants ($M = -\$292.78$, $SD = \$163.58$, $p = .05$, $d = .57$). Freeze probability participants ($M = -\$222.84$, $SD = \$142.01$) performed significantly better than did control participants, $p = .01$, $d = .76$. No other format differences reached significance. There was a significant main effect of forecast error, $F(1, 146) = 334.44$, $p < .01$, with significantly more EL in the high-error trials ($M = -\$412.12$, $SD = \$254.49$) than in the low-error trials ($M = -\$111.21$, $SD = \$69.91$), $d = 1.61$. There was also a significant interaction, $F(3, 146) = 4.10$, $p < .01$, suggesting a smaller decrement in performance due to high forecast error for uncertainty (frequency and probability) as compared to deterministic forecast formats (control and decision aid).

For the high forecast error first group there was a significant main effect for forecast error, $F(1, 130) = 319.62$, $p < .01$, with significantly worse performance in the high error trials ($M = -\$420.19$, $SD = \$194.56$) than in the low-error trials ($M = -\$169.48$, $SD = \$108.71$), $d = 1.59$. However, neither the main effect for forecast format or the interaction reached significance. To better understand the difference between the two orders, we compared them directly to one another. The primary impact of order appeared to be on the low-error trials. Performance in the low-error trials was significantly worse when the low forecast error block was presented after the high-error block ($M = -\$169.48$, $SD = \$108.71$) than when it was presented first ($M = -\$111.21$, $SD = \$69.91$), $t(282) = -5.43$, $p < .01$, $d = .64$, suggesting that the negative impact of bad forecasts endures even after the forecast improves. However the EL in high-error trials was similar across orders, $t(282) = -.30$, $p = .77$.

Table 2
Experiment 1: Mean Expected Loss Difference Scores (Participant EL–Optimal EL) on Low-Error-First Trials

Forecast format	Low-error trials *M* (*SD*)	High-error trials *M* (*SD*)	*Overall*
Control	−\$139.52 (\$80.03)	−\$521.41 (\$224.22)	*−\$330.47*
Decision Aid	−\$126.87 (\$75.09)	−\$458.68 (\$262.99)	*−\$292.78*
Freeze Frequency	−\$89.44 (\$55.92)	−\$319.96 (\$245.53)	*−\$204.70*
Freeze Probability	−\$90.79 (\$54.36)	−\$354.89 (\$240.48)	*−\$222.84*
Overall	*−\$111.21* (*\$69.91*)	*−\$412.12* (*\$254.49*)	*−\$262.70*

Note. Higher difference scores suggest better performance.

Trust. Our expectations for greater trust among those using uncertainty formats and in low-error forecasts were confirmed. There was a similar pattern in the daily and monthly trust ratings so only the monthly trust ratings will be discussed. A mixed-model ANOVA was conducted on the monthly trust ratings, with forecast format (freeze frequency, freeze probability, decision aid, and control) as the between-groups variable and forecast error (low and high) as the within-groups variable. For participants who encountered the low forecast error block first, there was a main effect for forecast format, $F(3, 146) = 7.84$, $p < .01$. Tukey's post hoc analyses revealed that freeze frequency participants ($M = 2.82$, $SD = .74$) had higher trust ratings than did either control participants ($M = 2.14$, $SD = .65$, $p < .01$, $d = .98$) or decision aid participants ($M = 2.21$, $SD = .69$, $p < .01$, $d = .85$). Freeze probability participants ($M = 2.58$, $SD = .71$) also had higher trust ratings than did control participants, $p = .04$, $d = .65$. No other format differences reached significance. There was a main effect for forecast error, $F(1, 146) = 99.52$, $p < .01$, with greater trust for low- ($M = 2.86$, $SD = .86$) as compared to high-error forecasts ($M = 2.02$, $SD = .94$), $d = .93$. The interaction failed to reach significance, suggesting that the trust advantage for uncertainty formats was maintained in the high-error trials (see Figure 1).

For participants who encountered the high forecast error block first, the pattern in trust ratings was similar. Participants rated high-error trials significantly lower ($M = 2.02$, $SD = .91$) than low error trials ($M = 2.36$, $SD = 1.09$), $F(1, 130) = 12.89$, $p < .01$, $d = .34$. Neither the main effect of format nor the interaction reached significance. As with EL, presenting high-error trials first had a negative impact on trust. Participants had significantly lower trust ratings in the low-error trials, $t(282) = 4.34$, $p < .01$, $d = .51$, when they were presented after the high forecast error block ($M = 2.36$, $SD = 1.09$) than when they were presented first ($M = 2.86$, $SD = .86$). Trust ratings in the high error trials were similar across orders, $t(282) = -.02$, $p = .98$.

Binary decision strategy. Thus, as predicted, there was a distinct advantage for uncertainty formats overall when the low forecast error trials were presented first. In addition, although

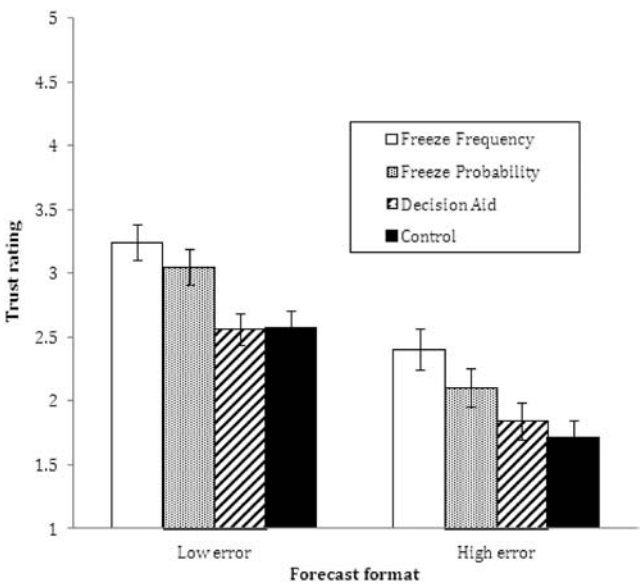

Figure 1. Mean monthly trust ratings by forecast format (Experiment 1). The rating scale ranged from 1, very little, to 5, very much. Error bars represent +/− 1 standard error.

forecast error had a negative impact on budget balance, EL and trust, the negative impact was reduced for participants with uncertainty information. To better understand the decision strategies that lead to these advantages for the uncertainty formats, we examined the binary decisions more closely in the low error first order. For these analyses the uncertainty groups were combined because there were no significant differences between them in the previous analyses. Likewise the control and decision aid conditions were combined to form a deterministic condition because they were not significantly different in the previous analyses and because neither included an uncertainty estimate.

In order to determine whether participants made different decisions in the ranges of PoF above and below the economically rational threshold, we calculated the proportion of salt decisions in each of the six probability ranges and then calculated the average proportion above and below the 17% threshold for each participant (see Figure 2). Then, we conducted a mixed-model ANOVA on mean proportion of "salt" responses per participant, with PoF range (above and below 17%) as the within-groups variable and forecast format (uncertainty and deterministic) as the between-groups variable. There was a significant main effect for PoF range, $F(1, 148) = 1270.00$, $p < .01$. Participants salted more above the 17% threshold ($M = .74$, $SD = .17$) than below it ($M = .21$, $SD = .20$), $d = 2.86$. There was also a significant interaction, $F(1, 148) = 34.09$, $p < .01$, suggesting greater differentiation in salt decisions among those with uncertainty formats. Those with uncertainty forecasts 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 ($M = .70$, $SD = .18$).

Next we asked how strategy changed when forecast error increased, comparing decisions in high- and low-error trials directly. Recall that the two blocks had the same set of deterministic forecasts, ranging from 32 °F to 35 °F. However, the probability of discrepant observed temperatures increased in the high-error trials because of the increase in variance. As a result, 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, we focus here on the 22 trials in which the appropriate course of action was the same in both the high- and low-error trials, to salt (PoF $\geq$ 17%). A mixed-model ANOVA was conducted on the proportion of participants salting with forecast format (uncertainty and deterministic) as the between-groups variable and forecast error (low and high) as the within-groups variable. There was 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 ($M = .61$, $SD = .20$), $d = .36$. 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 ($M = .63$, $SD = .18$). In sum, there was a decrease in decisions to salt among those relying on deterministic forecasts alone, suggesting an increase in risk-seeking. However, for those with uncertainty forecasts, the opposite pattern was observed. They continued to do better, taking action more often despite the increase in forecast error.

Decision errors. Finally, we examined decision errors defined in terms of the economically optimal strategy. Here we examine the four conditions separately again. There were two kinds of errors: Failing to salt when PoF was greater than 17%, incurring more risk than was optimal (risk-seeking), and salting when the PoF was less than 17%, spending more than was necessary to protect against risk (risk-averse). In order to determine which kind of error was prevalent, we focused on the range of

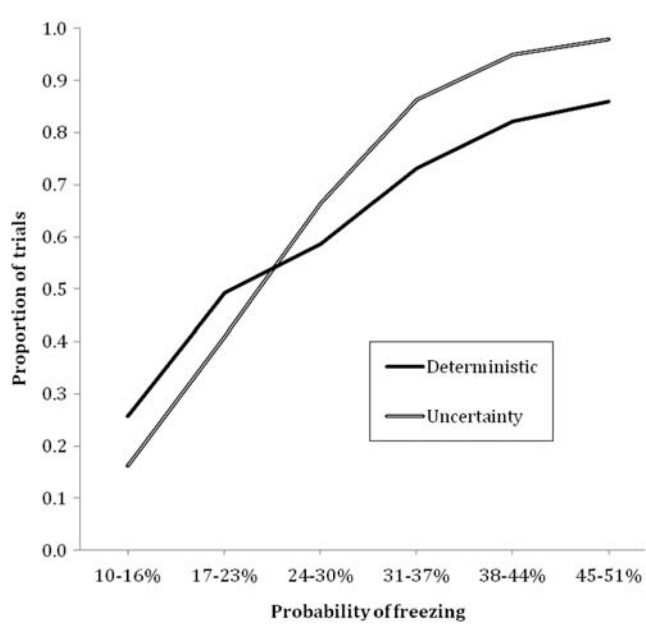

Figure 2. Proportion of trials on which participants (control vs. experimental) salted plotted over probability of freezing range categories (Experiment 1).

probabilities of freezing between 10% to 23%, in which there was an equal number of trials above and below 17% PoF. A decision *not* to salt was coded as an error above 17% and decision to salt was coded as an error below 17%. Then, a mixed-model ANOVA was conducted on percent "errors," with error type (risk-seeking and risk-averse) as the within-groups variable and forecast format (freeze frequency, freeze probability, decision aid, and control) as the between-groups variable. There was a significant main effect for error type, $F(1, 146) = 113.41$, $p < .01$. Participants made more risk-seeking ($M = .55$, $SD = .25$) than risk-averse errors ($M = .21$, $SD = .20$), $d = 1.50$. There was also a significant interaction, $F(3, 146) = 5.06$, $p < .01$. Participants with uncertainty information tended to make fewer risk-averse but more risk-seeking errors than did participants with deterministic forecasts (see Table 3).

Thus, Experiment 1 demonstrated that people made better decisions overall when they had uncertainty information than when they did not. This included, for the first time, evidence that decisions based on uncertainty forecasts are superior to decisions based on advice about the optimal course of action. Furthermore, Experiment 1 demonstrated the extensive negative impact of forecast error on both performance and trust. Participants were clearly aware of the increased forecast error, rating the high-error trials as significantly less trustworthy than the low-error trials. Moreover, the negative impact of forecast error was long lasting. When high forecast error was encountered in the first block of trials, the reduction in both performance and trust continued into the low forecast error trials, suggesting that trust once lost is difficult to regain even when the forecast becomes more reliable, at least over the 60 trials observed here.

In addition, when we examined decisions among participants encountering low forecast error first, we noted that those with only deterministic forecasts became more risk-seeking in the high-error trials. This suggests that when precautionary action involves a cost and forecast error increases, people are reluctant to act. Moreover, this experiment demonstrated for the first time that reliable uncertainty information attenuates this negative effect, allowing participants to continue to take precautionary action in the face of increased forecast error. This pattern was also reflected in trust ratings, with higher ratings among those using uncertainty formats in both the low- and high-error trials. Only when the probability of freezing was just above the 17% threshold action did participants using uncertainty forecasts make more errors than those using deterministic forecasts. Contrary to our expectations, the frequency expression did not help in this range. Perhaps a stronger manipulation to increase the salience of the cost is required. This is one of the questions asked in Experiment 2.

Although participants in the deterministic conditions performed better in the range between 17% and 23%, they salted more often overall in the low probabilities both above and below (10%−17%) the threshold than did those with uncertainty forecasts. Perhaps, because they were unsure of the outcome, they salted some of the time regardless of the forecast in an attempt to cover all possibilities. Although this strategy led to an advantage in the PoF range immediately above the 17% threshold, the cumulative effect, over the entire range of probabilities, was negative.

Surprisingly, giving participants specific decision recommendations did not help as compared to the control condition, by any measure of performance or trust. In fact, participants made decisions in accord with the advice on average only 60% of the time, not significantly more often than did control participants (56%) who did not see the advice. This may have been because the decision aid advised participants to salt whenever the PoF was greater than 17%, when freezing temperatures were observed on only 37 out of 102 trials (36.3%). As such, following the advice often meant spending $1,000 to salt when observed temperatures were above freezing, making salting unnecessary. This may have seemed like a waste of money from participants' perspective. They may not have understood that the advantage for following the advice was realized over several trials, because the penalty for not salting was so great that it offset the cost of unnecessary salting up to six times. Perhaps if this were explained to participants, they would be more likely to follow the advice. This hypothesis was also tested in Experiment 2.

Table 3
Experiment 1: Proportion of Trials on Which Participants Made a Risk-Seeking or Risk-Averse Error

	Risk-averse error	Risk-seeking error	
Forecast Format	*M* (*SD*)	*M* (*SD*)	*Overall*
Control	.30 (.23)	51 (.24)	*.40*
Decision Aid	.22 (.16)	.51 (.23)	*.36*
Deterministic	*.26 (.20)*	*.51 (.23)*	*.38*
Freeze Frequency	.21 (.23)	.54 (.27)	*.38*
Freeze Probability	.10 (.10)	.65 (.22)	*.37*
Uncertainty	*.16 (.18)*	*.60 (.25)*	*.38*
Overall	*.21 (.20)*	*.55 (.25)*	*.38*

Note. A risk-averse error was applying salt treatment when the probability of freezing was <17%. A risk-seeking error was withholding salt treatment when the probability of freezing was ≥17%.

Experiment 2

Experiment 2 employed the same basic decision task and procedure as Experiment 1 to test three new forecast formats in an attempt to improve performance. One was an uncertainty forecast that emphasized the consequences of inaction by describing the $6,000 fine on each forecast. We hypothesized that this would encourage people to take action even when the probability of freezing was low because it would be clear that the potential loss was great. Because this manipulation made the most sense in context of the frequency phrase, only the freeze frequency format was tested in Experiment 2.

The other two new forecast formats were designed to convince participants to follow the advice of the decision aid. Both included more detailed and slightly different explanations of the economic rationale and acknowledged that salting would sometimes be required when the likelihood of freezing was low (see Appendix).

Method

Participants. A total of 255 University of Washington psychology students (47.84% female) participated for course credit and the chance to earn prize money. Mean age was 19.13 years (range16−29). None of them participated in Experiment 1.

Procedure. The computer-administered procedure was identical to that in Experiment 1 with two exceptions. In Experiment 2,

to simplify the payment schedule, participants received a virtual monthly budget of \$30,000 for each of four hypothetical months and were paid \$2 for every \$1,000 remaining in their budget at the end of the experiment. Forecast error was not manipulated in Experiment 2, although there were four blocks of trials as with Experiment 1. Here, the third and fourth blocks were included as filler trials so that analyzed trials would not include any in which participants had a negative balance without pending budget installments.

Design. Experiment 2 manipulated a single factor, forecast format, between participants, who were randomly assigned to one of six conditions, all including the deterministic forecast. The control, freeze frequency and decision aid conditions were identical to those tested in Experiment 1. The "freeze frequency consequences" condition also included the penalty for not salting when freezing temperatures were observed (e.g., "22 out of 100 days like this the temperature will be equal to or less than 32 degrees and you will be penalized \$6,000 if you do not salt"). The "decision aid *calculation*" condition included the mathematical calculation used to generate the advice. The "decision aid *explanation*" condition described the probabilistic nature of the decision recommendation and the long-run benefit of following it (see Appendix). The dependent variables were expected loss (EL), trust rating and binary salt decision.

Stimuli. The same 60 low-error forecasts and observations used in Experiment 1 were again used in Experiment 2.

Results

Eight participants who were at least two standard deviations above the mean standard error in temperature difference score were omitted from the analyses below, leaving at total of 247 participants.

Expected loss. EL difference was calculated for each participant using the same procedure as described in Experiment 1, and subtracted from the optimal EL (see Table 4). A one-way ANOVA on EL difference scores with forecast format (control, freeze frequency, freeze frequency consequences, decision aid, decision aid calculation, decision aid explanation) as the independent variable revealed a significant main effect for format, $F(5, 246) = 3.65$, $p < .01$. Tukey's post hoc analyses indicated significantly better performance among freeze frequency ($M = -\$96.42$, $SD = \$70.14$, $p = .02$, $d = .67$) and freeze frequency consequences participants ($M = -\$98.76$, $SD = \$54.44$, $p = .02$, $d = .69$) than among control participants ($M = -\$154.27$, $SD = \$100.11$). No other differences reached significance.

Trust. An identical one-way ANOVA on trust ratings revealed a significant main effect for format, $F(5, 246) = 4.71$, $p < .001$. Tukey's post hoc analysis indicated significantly greater trust among freeze frequency ($M = 2.65$, $SD = .70$, $p = < .01$, $d = .88$) and freeze frequency consequences participants ($M = 2.79$, $SD = .70$, $p < .01$, $d = 1.09$) than among control participants ($M = 2.05$, $SD = .66$). See Figure 3. No other differences reached significance.

Binary decision strategy. Because of the lack of significant differences between the uncertainty conditions and between the control and decision aid conditions, they were consolidated into two groups, uncertainty and deterministic, to examine decision strategy. A mixed-model ANOVA on proportion of "salt" decisions, with PoF range (above and below 17%) as the within-groups variable and forecast format (uncertainty and deterministic) as the between-groups variable revealed a significant main effect for PoF range, $F(1, 245) = 1666.16$, $p < .01$. Participants salted more often above the 17% threshold ($M = .73$, $SD = .15$) than below it ($M = .22$, $SD = .19$), $d = 2.98$. There was also a significant interaction, $F(1, 245) = 26.30$, $p < .01$, suggesting greater differentiation in salt decisions among those with uncertainty formats (see Figure 4). They salted less below the 17% threshold ($M = .16$, $SD = .16$) than did those with deterministic forecasts ($M = .25$, $SD = .20$), and more above the 17% threshold ($M = .76$, $SD = .12$) than did those with deterministic forecasts ($M = .72$, $SD = .16$).

Decision errors. Finally, we examined decision errors in the range of PoF between 10 and 23% in which there were equal opportunities for risk-seeking and risk-averse errors. For this analysis all of the original 6 conditions were tested separately. A mixed-model ANOVA was conducted on percent "errors" with error type (risk-seeking and risk-averse) as the within-groups variable and forecast format as the between-groups variable. There was a significant main effect for error type, $F(1, 241) = 163.44$, $p < .01$. Participants made more risk-seeking ($M = .52$, $SD = .23$) than risk-averse errors ($M = .22$, $SD = .19$), $d = 1.42$. There was also a significant interaction, $F(5, 241) = 4.4$, $p < .01$. Participants in the uncertainty conditions made fewer risk-averse but more risk-seeking errors. Importantly, there was a significant main effect for forecast format, $F(5, 241) = 2.99$, $p = .01$. Tukey's post hoc tests revealed that those in the decision aid ($M = .35$, $SD = .08$, $p = .03$, $d = .62$) and decision aid calculation ($M = .34$, $SD = .09$, $p = .02$, $d = .70$) conditions made significantly fewer errors than did those in the control condition ($M = .41$, $SD = .11$). No other post hoc differences reached significance. Thus, although there was no advantage for explicit advice over the entire range of PoF, there was an advantage in the low probability range around the threshold for action (see Table 5).

Experiment 2 replicated the major advantages for uncertainty formats observed in Experiment 1 in terms of EL, trust and decision strategy. However, reminding participants of the consequences of inaction on each trial did not improve performance over the advantage for uncertainty information alone. We are reluctant to draw conclusions from this null result because it may be due to an aspect of the experimental task that is dissimilar to actual weather-related decisions. Because one trial followed immediately after the next, participants in Experiment 2 may have been sufficiently aware of the consequences of inaction that the reminder

Table 4
Experiment 2: Mean Expected Loss Difference Scores (Participant EL–Optimal EL)

Forecast format	*M* (*SD*)
Control	−\$154.27 (\$100.11)
Decision Aid	−\$141.95 (\$80.48)
Decision Aid Calculation	−\$134.77 (\$75.56)
Decision Aid Explanation	−\$134.34 (\$89.05)
Freeze Frequency	−\$96.42 (\$70.14)
Freeze Frequency Consequences	−\$98.76 (\$54.44)

Note. Higher difference scores suggest better performance.

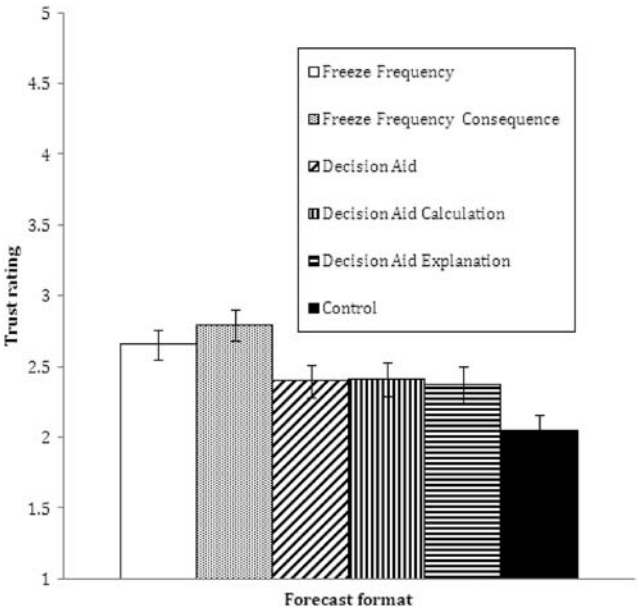

Figure 3. Mean monthly trust ratings by forecast format (Experiment 2). The rating scale ranged from 1, very little, to 5, very much. Error bars represent +/− 1 standard error.

was not necessary. There are comparatively long intervals between weather warnings during which more forgetting may occur, possibly making reminders useful. Perhaps more surprising was the fact that the additional explanations added to the decision aid did *not* increase compliance with it overall. Again, participants receiving the advice did not make decisions in accord with it significantly more often than did those in the control condition who did not receive it.

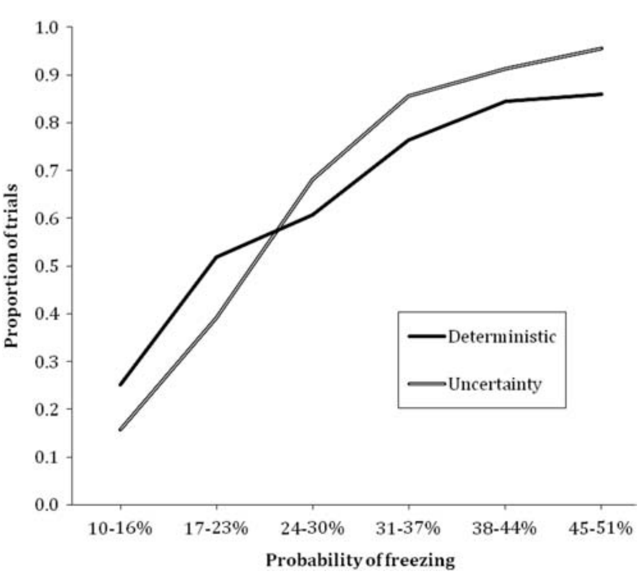

Figure 4. Proportion of trials on which participants (deterministic vs. uncertainty) salted plotted over probability of freezing range categories (Experiment 2).

Table 5
Experiment 2: Proportion of Trials on Which Participants Made a Risk-Seeking or Risk-Averse Error

Forecast Format	Risk-averse error *M* (*SD*)	Risk-seeking error *M* (*SD*)	*Overall*
Control	.29 (.21)	.52 (.23)	*.41*
Decision Aid	.27 (.18)	.43 (.17)	*.35*
Decision Aid Calculation	.20 (.18)	.48 (.23)	*.34*
Decision Aid Explanation	.24 (.22)	.50 (.26)	*.37*
Freeze Frequency	.14 (.17)	.62 (.24)	*.38*
Freeze Frequency Consequence	.18 (.15)	.59 (.21)	*.39*
Overall	*.22 (.19)*	*.52 (.23)*	*.37*

Note. A risk-averse error was applying salt treatment when the probability of freezing was <17%. A risk-seeking error was withholding salt treatment when the probability of freezing was ≥17%.

We did, however, detect an advantage for the decision aid in trials with low probabilities surrounding the optimal threshold. Thus in order to maintain performance over the full range of probabilities the solution may be to combine the decision advice with an uncertainty estimate. The decision aid may help when the probabilities are between 17% and 23% because without it, the possibility of freezing may seem too remote to warrant salting for those unaware of the principles of expected value. Thus, explicit advice may be necessary to prompt action in these situations. For probabilities over the full range, the advice alone, yielding numerous false alarms may seem inaccurate and untrustworthy until an estimate is added acknowledging uncertainty. Indeed, this combination is the most ecologically valid, as emergency managers are required to give evacuation notices regardless of whether uncertainty estimates are added. This combined format was tested in Experiment 3.

Experiment 3

The purpose of Experiment 3 was to test whether people would make better decisions overall if they had forecasts that included both uncertainty information and the decision aid guidance. To test this hypothesis, the decision aid, identical to that in Experiment 1, was accompanied in one condition by the probability of freezing and in another by the frequency of freezing.

In addition, to replicate the advantage for uncertainty formats when forecast error increased, we included the high forecast error trials tested in Experiment 1. However, in Experiment 3, we presented the high-error trials in the second half for all conditions and we restarted participants' balances at the beginning of the high-error block to increase motivation for these trials and provide a stronger test of the hypothesis that increased error alone is responsible for decline in performance in the high-error trials.

Method

Participants. A total of 178 University of Washington psychology students (48.31% female) participated for course credit and the chance to earn prize money. Mean age was 19.48 years (range 18–28). None of them participated in Experiment 1 or 2.

Procedure. The procedure was identical to that of Experiment 1 with the budget and payment schedule in Experiment 2. However, in Experiment 3 participants' budgets were reset at the beginning of the third month so that they began the high-error trials with \$30,000 regardless of their balance at the end of the low-error block. At the end of the experiment, participants' balances from the end of the second and fourth months were summed, and participants with total ending balances above zero were paid \$2 for every \$1,000 in their final balance.

Design. A 3 (between) × 2 (within) mixed-model design was used. The between-groups variable was forecast format. Participants were randomly assigned to one of three forecast formats, all of which included the deterministic forecast. In the control condition, only the deterministic forecast was provided. The two experimental conditions included decision aid and either the frequency of freezing or the probability of freezing, each expressed as it had been in the two previous experiments.

The within-groups variable was the amount of error in the deterministic forecast. The first block of 60 trials had a mean standard error of 3.26 (low-error trials) and the second had a mean standard error of 6.53 (high-error trials). The dependent variables were expected loss (EL), trust rating and binary salting decision.

Stimuli. The forecasts and observations were identical to those used in Experiment 1.

Results

Eight participants who were at least two standard deviations above the mean standard error in temperature difference scores were omitted, leaving a total of 170 participants. As with the previous two experiments we examined only comparable trials in which participants could expect a new budget installment.

Expected loss. EL for months 1 and 3 was calculated for each participant, subtracting the optimal EL (see Table 6). A mixed-model ANOVA on EL difference scores, with forecast error (low and high) as the within-groups variable and forecast format (control, decision aid frequency, decision aid probability) as the between-groups variable, revealed a significant main effect for forecast error, $F(1, 167) = 426.02$, $p < .01$, with better performance in low-error trials ($M = -\$115.45$, $SD = \$77.77$) than in high-error trials ($M = -\$452.57$, $SD = \$253.17$), $d = 1.80$. There was a significant main effect for forecast format, $F(2, 167) = 13.83$, $p < .01$. A Tukey's post hoc analysis indicated better performance among decision aid frequency ($M = -\$273.06$, $SD = \$143.52$, $d = .62$) and decision aid probability ($M = -\$219.34$, $SD = \$147.26$, $d = .99$) participants than among control participants ($M = -\$357.56$, $SD = \$130.09$), $ps < .01$. There was also a significant interaction, $F(2, 167) = 9.17$, $p < .01$, indicating less decline in performance from low- to high-error trials for decision aid frequency ($M = -\$327.00$, $SD = \$211.26$) and decision aid probability participants ($M = -\$255.36$, $SD = \$232.19$) than for the control participants ($M = -\$426.29$, $SD = \$192.46$).

Trust. An identical mixed-model ANOVA on mean monthly trust rating showed a significant main effect for forecast error $F(1, 167) = 64.03$, $p < .01$, with less trust on high-error trials ($M = 2.08$, $SD = .98$) than on low-error trials ($M = 2.73$, $SD = .88$), $d = .70$. There was also a significant forecast format effect, $F(2, 167) = 3.48$, $p = .03$, with higher trust ratings among the experimental groups. A Tukey's post hoc analysis revealed significantly higher trust ratings among the decision aid frequency participants ($M = 2.61$, $SD = .76$) than among control participants ($M = 2.26$, $SD = .73$), $p = .04$, $d = .47$. See Figure 5.

Binary decision strategy. Again, the two experimental conditions (decision aid frequency and decision aid probability) were not significantly different from one another so they were combined to examine decision strategy. A mixed-model ANOVA on proportion of "salt" decisions in the low error trials, with PoF range (above and below 17%) as the within-groups variable and forecast format (experimental and control) as the between-groups variables, revealed a significant main effect for PoF range, $F(1, 168) = 834.24$, $p < .01$. Participants salted more often above the 17% threshold ($M = .72$, $SD = .16$) than below it ($M = .19$, $SD = .19$), $d = 3.02$. There was a significant interaction, $F(2, 168) = 49.24$, $p < .01$, indicating that experimental participants salted less often below the 17% threshold ($M = .15$, $SD = .17$) than did the control participants ($M = .27$, $SD = .19$) and more often above it ($M = .76$, $SD = .15$) than did the control participants ($M = .64$, $SD = .17$). This pattern is shown in Figure 6, which also includes the decision aid condition from Experiment 1 to illustrate the overall advantage for the combined formats when the uncertainty estimate is added.

Next we asked how strategy changed when forecast error increased. As with Experiment 1, we compared the proportion of salt decisions on the matching high- and low-error trials for which the appropriate course of action was to salt. A mixed-model ANOVA was conducted on the proportion of decisions to salt with forecast format (experimental and control) as the between-groups variable and forecast error (low and high) as the within-groups variable. There was a significant main effect for forecast error, $F(1, 168) = 12.76$, $p < .01$, showing that participants salted more in the

Table 6
Experiment 3: Mean Expected Loss Difference Scores (Participant EL–Optimal EL)

Forecast format	Low-error trials *M* (*SD*)	High-error trials *M* (*SD*)	*Overall*
Control	−\$144.41 (\$78.05)	−\$570.70 (\$215.11)	*−\$357.56*
Decision Aid Frequency	−\$109.56 (\$77.89)	−\$436.56 (\$239.67)	*−\$273.06*
Decision Aid Probability	−\$91.66 (\$68.63)	−\$347.02 (\$256.16)	*−\$219.34*
Overall	*−\$115.45 (\$77.77)*	*−\$452.57 (\$253.17)*	*−\$283.32*

Note. Higher difference scores suggest better performance.

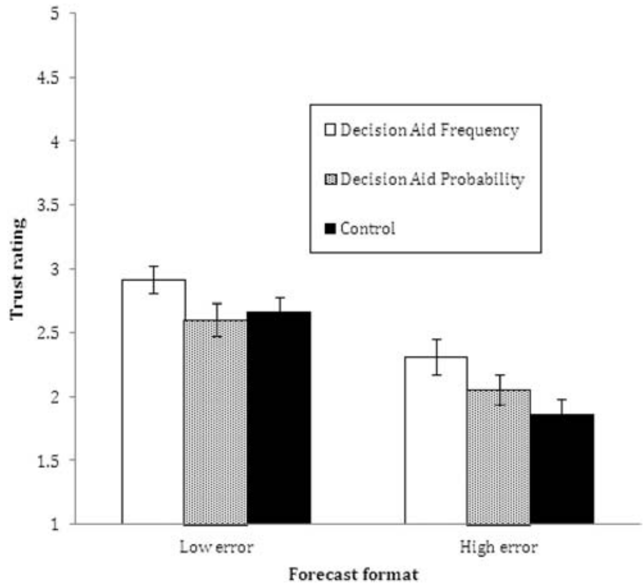

Figure 5. Mean monthly trust ratings by forecast format (Experiment 3). The rating scale ranged from 1, very little, to 5, very much. Error bars represent +/− 1 standard error.

low-error trials (M = .65, SD = .17) than in the high-error trials (M = .61, SD = .24), d = .19. There was a significant main effect for forecast format, $F(1, 168) = 12.50$, $p < .01$. Participants in the experimental conditions salted more often (M = .66, SD = .18) than did participants in the control condition (M = .56, SD = .17), d = .57. Moreover, there was a significant forecast error by format interaction, $F(1, 168) = 18.03$, $p < .01$. Participants in the control condition salted less often in the high- (M = .50, SD = .21) as

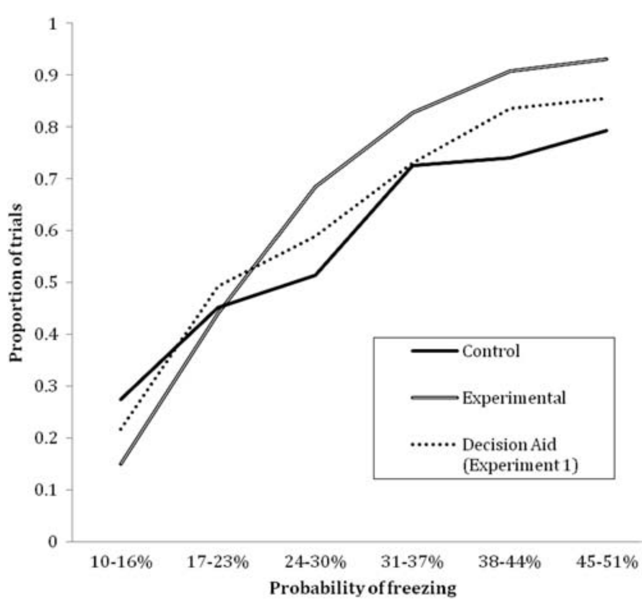

Figure 6. Proportion of trials on which participants (control vs. experimental) salted plotted over probability of freezing range categories (Experiment 3), including data from decision aid participants (Experiment 1).

compared to the low- (M = .63, SD = .17) error trials, while those in the experimental condition salted approximately equally often in the low- (M = .66, SD = .17) and the high- (M = .67, SD = .24) error trials. Thus, the change in strategy with forecast error that was detected in Experiment 1 was replicated here in Experiment 3. When forecast error was high, people with conventional deterministic forecasts salted less often when it was economically rational to do so, becoming more risk-seeking. However, those in the experimental conditions continued to make better decisions, despite the increase in forecast error.

To this point, results from the two new forecast formats combining the decision advice with uncertainty forecasts have replicated the advantages for uncertainty formats found in Experiments 1 and 2. Experimental participants had significantly lower expected loss and greater trust in both the low and high forecast error trials. In addition there was less decline in performance with the increase in forecast error among those in the experimental conditions. This suggests that adding explicit advice to the uncertainty forecast did not diminish any of the previously observed advantages.

Decision errors. Finally, we examined decision errors in the range of PoF between 10 and 23% in which there were equal opportunities for risk-seeking and risk-averse errors. Here we tested conditions separately to determine whether adding explicit advice to each of the uncertainty formats improved performance in these situations. A mixed-model ANOVA was conducted on percent "errors" with error type (risk-seeking and risk-averse) as the within-groups variable and forecast format (decision aid frequency, decision aid probability, and control) as the between-groups variable. There was a significant main effect for error type, $F(1, 167) = 160.69$, $p < .01$. Participants made more risk-seeking errors (M = .56, SD = .23) than risk-averse errors (M = .19, SD = .19), d = 1.75. There was a significant main effect for format, $F(2, 167) = 8.86$, $p < .01$. Tukey's post hoc analyses indicated that there were significantly smaller proportions of errors in both the decision aid frequency (M = .37, SD = .09, p = .02, d = .47) and decision aid probability (M = .34, SD = .10, $p < .01$, d = .77) conditions than in the control condition (M = .41, SD = .08), replicating the advantage for the decision aid conditions observed in Experiment 2. There was also a significant interaction, $F(2, 167) = 4.22$, p = .02, suggesting that there was greater differentiation among formats in the trials below as compared to above the threshold (see Table 7).

Table 7
Experiment 3: Proportion of Trials on Which Participants Made a Risk-Seeking or Risk-Averse Error

	Risk-averse error	Risk-seeking error	
Forecast format	*M* (*SD*)	*M* (*SD*)	*Overall*
Control	.27 (.19)	.55 (.21)	*.41*
Decision Aid Frequency	.13 (.16)	.60 (.23)	*.37*
Decision Aid Probability	.17 (.19)	.51 (.25)	*.34*
Overall	*.19* (*.19*)	*.56* (*.23*)	*.37*

Note. A risk-averse error was applying salt treatment when the probability of freezing was <17%. A risk-seeking error was withholding salt treatment when the probability of freezing was ≥17%.

Participants in the experimental conditions using the decision aid accompanied by an uncertainty estimate clearly did better by all standards in Experiment 3. In order to determine whether this was due to increased compliance with the advice, we examined the proportion of trials on which participants followed it. Here for the first time, that proportion was significantly greater in both the decision aid probability ($M = 68\%$, $SD = 15.26\%$, $p < .01$, $d = 1.02$) and decision aid frequency ($M = 61.82\%$, $SD = 14.42\%$, $p < .01$, $d = .57$) conditions than in the control condition ($M = 54.28\%$, $SD = 12.11\%$). Thus, in addition to replicating all of the advantages for uncertainty forecasts, Experiment 3 demonstrated a clear advantage for the combination of advice and uncertainty estimate in the sensitive low probability of freezing range as well as in compliance with the advice.

General Discussion

This group of experiments provides strong evidence that non-experts make better precautionary decisions under uncertainty when they have an explicit description of the risk involved, replicating the results of the handful of studies that have made the crucial comparison between decisions made with and without uncertainty estimates. Here, people made better decisions when the adverse event was likely, taking action more often than did those with deterministic forecasts alone. They also made better decisions when it was unlikely, withholding action more often than did those with deterministic forecasts. It is important to note that although participants did not make optimal decisions from the perspective of economic models of rational choice, their decisions were closer to that standard than when uncertainty estimates were not provided. Moreover, the effects for these advantages for all of the major variables including budget, expected loss, and trust ranged from medium to large in size, suggesting that they are of practical as well as theoretical significance.

Critically, these experiments demonstrated for the first time that those with uncertainty forecasts performed significantly better, by all measures, than did those who were given concrete advice about the economically optimal course of action. It is now clear that telling people what to do without giving them the relevant uncertainty information is not compelling, even when the advice itself is good. In the first two experiments reported here, decisions based on advice describing the optimal choice, were no better than those made in the control condition. Surprisingly, overall performance was not improved by either explaining the calculations or by explaining the long run advantage of following the recommendations. The only thing that helped was adding an uncertainty estimate in Experiment 3, providing participants the information to judge for themselves.

In addition, these experiments provided evidence that uncertainty forecasts are advantageous in situations in which error in the deterministic forecast increases. One of the major claims made about uncertainty forecasts heretofore untested is that they promote trust, especially when the single-value forecast becomes less accurate. Inevitably, weather forecasts, like forecasts of all kinds, encounter situations that are less predictable and forecast accuracy decreases. Here, for the first time, we have evidence to support the claim that forecasts that acknowledge the uncertainty seem less “wrong” when the single-value forecast fails to verify. Although all participants encountered the same sequence of low temperature forecasts and observations with the same increase in error, those using uncertainty forecasts rated them as significantly more trustworthy than did those using deterministic forecasts. Moreover, people with uncertainty forecasts performed better by every measure (EL, budget, binary decisions) in the high-error trials after trust had been established in the low-error trials than did those with only the deterministic forecast. The advantage here was likely twofold. People with uncertainty forecasts continued to trust the forecast and use it to make their decision. As well, the uncertainty estimate, which continued to be well calibrated, constituted additional useful information.

These experiments also tackled a critical practical issue, the failure to take action in the face of adverse weather events. According to FEMA, estimated compliance with mandatory evacuation is notoriously low, ranging between 40 and 70% (Hurricane Forecast Improvement Project, 2010). Many residents choose to stay in their homes even when they are told that it is dangerous to do so and even when transportation has been provided (Morss & Hayden, 2010). Survey research suggests that their decisions are influenced by the cost of evacuation itself, including the dangers of being out on the highway, health risks, and leaving their home unprotected (Baker, 1991; Dow & Cutter, 2000; Morss & Hayden, 2010;Smith & McCarty, 2009). In the decision task used here, we replicated this problem. The most common error among our participants was a failure to take precautionary action when they should have protected themselves against the penalty for missing a freeze event, a tendency that *increased* with error in the forecast.

Why do people fail to act in situations when precautionary action is appropriate? Our results suggest that to a certain extent it is because they do not understand the risk. Participants with calibrated uncertainty information made fewer risk-seeking choices above the optimal threshold for action than did participants using deterministic forecasts and the decision support advice. This suggests that people fail to take action in some cases, even when they are told to do so, because they do not have an accurate understanding of the risk. When participants were left to make their own risk assessment they often decided, despite the advice to the contrary, that action was not warranted. Indeed, their subjective estimates of amount of uncertainty appeared to be greater across the board than the actual uncertainty involved. Those in the deterministic conditions tended to withhold salt more often when the probability of freezing was high and salt more often when the probability of freezing was low, suggesting that they were trying to cover a wider range of possibilities. Their choices were not as clearly differentiated across the range of weather situations, as were the choices of those with uncertainty forecasts. This suggests that explicit uncertainty estimates improved decisions by providing the narrower range of possibilities to anticipate for any given situation, reducing both kinds of error.

The results reported here also suggest that failing to act in complex realistic situations is due in part to distrust in the forecast based on prior experience. Participants without uncertainty estimates reacted to increased forecast error by salting less often. Recall that the low- and high-error temperature forecasts were identical to one another; it was the outcome, the observed temperature, that differed resulting in double the unbiased error in the high-error trials. Thus, faced with an identical set of temperature forecasts, participants in the deterministic conditions took action significantly less often in the high-error trials. This suggests that

when the prediction is unreliable people may simply do nothing rather than take action that may constitute a mistake. Unfortunately, these results also suggest that the effects of erroneous forecasts persist, even after the forecast itself recovers accuracy. In Experiment 1, when high-error forecasts were presented first, performance decrements endured throughout the low-error trials.

The reluctance to act in the face of forecast error observed here may explain much of the noncompliance with weather warnings. If people perceive the forecast to be in error, they may be more likely do nothing, staying in their homes even when it is not safe to do so. It is important to note, however, that these problems were attenuated in the experiments reported here by uncertainty forecasts, as reflected in budget, EL and trust ratings in high-error trials. This suggests that forecasts that acknowledge the uncertainty are more believable to people and continue to provide useful information even when the single-value forecast becomes less reliable.

However, explicit uncertainty forecasts alone did not help in the lowest probability ranges above the rational threshold for action (PoF = 17–23%) in Experiments 1 and 2. The reluctance to take precautionary action in this range may indeed be true risk-seeking in a psychological sense. Here, there was no improvement with explicit uncertainty estimates suggesting that even when participants were aware of the amount of risk, they determined that it was worth taking.

Thus, failure to act in situations outside of the laboratory is probably a combination of these and other factors. However, giving explicit warning advice alone does not reduce this tendency. People understand that forecasts involve uncertainty, which makes predictions of "certain death" implausible to them. Only when the advice was combined with information that acknowledged and quantified the uncertainty was the error significantly reduced. This combination forecast may successfully convey the idea that when the potential loss is taken into account, the expense of precautionary action is warranted despite the low probability. This kind of forecast, making the reasoning transparent and the relevant information available to end-users, appears to be the most convincing.

These results have implications for a broad range of domains in which decisions are made under uncertainty by nonexperts. They suggest that explicit uncertainty estimates allow people to make better decisions when the circumstances are sufficiently motivating. Regardless of background or training, people understand that they live in an uncertain world: medical treatments have risks, the stock market is fickle, and they may or may not be in the path of severe weather. Thus, although uncertainty expressions are psychologically challenging, having a reliable uncertainty estimate can improve both the quality of the decisions as well as trust in the information source itself.

References

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

Barnes, L. R., Gruntfest, E. C., Hayden, M. H., Schultz, D. M., & Benight, C. (2007). False alarms and close calls: A conceptual model of warning accuracy. *Weather and Forecasting, 22,* 1140–1147. doi:10.1175/WAF1031.1

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society, 22,* 23–36.

Betts, S. C., & Taran, Z. (2005). Brand as a reliability reference point: A test of prospect theory in the used car market. *Journal of the Academy of Business and Economics, 5,* 34–38.

Blendon, R. J. (2008). High-risk area hurricane survey. Retrieved from Harvard School of Public Health, http://www.hsph.harvard.edu/hurricane/topline.doc

Budescu, D. V., Broomel, S., & Por, H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science, 20,* 299–308. doi:10.1111/j.1467-9280.2009.02284.x

Cutter, S. L., & Smith, M. M. (2009). Fleeing from the hurricane's wrath: Evacuation and the two Americas. *Environment, 51,* 26–36. doi:10.3200/ENVT.51.2.26-36

Dow, K., & Cutter, S. L. (2000). Public orders and personal opinions: Household strategies for hurricane risk assessments. *Environmental Hazards, 2,* 143–155. doi:10.3763/ehaz.2000.0220

Gneiting. T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science, 310,* 248–249. doi:10.1126/science.1115255

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 45th Annual Meeting* (pp. 1699–1703). Santa Monica, CA: Human Factors and Ergonomics Society.

Haas, M. L. (2001). Prospect theory and the Cuban Missile Crisis. *International Studies Quarterly, 45,* 241–270. doi:10.1111/0020-8833.00190

Hurricane Forecast Improvement Project (HFIP) Product Development. (2010, April). Workshop conducted at the National Center for Atmospheric Research (NCAR), Boulder, CO.

Joslyn, S., Pak, K., Jones, D., Pyles, J., & Hunt, E. (2007). The Effect of probabilistic information on threshold forecasts. *Weather and Forecasting, 22,* 804–812. doi:10.1175/WAF1020.1

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications, 17,* 180–195. doi:10.1002/met.190

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Marketing Science, 22,* 393–410. doi:10.1287/mksc.22.3.393.17737

Kahneman, D., & Tverksy, A. (1984). Choices, values, and frames. *American Psychologist, 39,* 341–350. doi:10.1037/0003-066X.39.4.341

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–292. doi:10.2307/1914185

Kerstholt, Jose, H. (1995). Decision making in a dynamic situation: The effect of false alarms and time pressure. *Journal of Behavioral Decision Making, 8,* 181–200. doi:10.1002/bdm.3960080304

Koehler, J. J. (2001). When are people persuaded by DNA match statistics? *Law and Human Behavior, 25,* 493–513. doi:10.1023/A:1012892815916

Lees, M. N., & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics, 50,* 1264–1286. doi:10.1080/00140130701318749

McKinley, J. C., & Urbina, I. (2008, September 13). A million flee as huge storm hits Texas coast. *The New York Times.* Retrieved http://www.nytimes.com

Morss, R. E., & Hayden, M. H. (2010). Storm surge and "certain death": Interviews with Texas coastal residents following Hurricane Ike. *Weather, Climate, & Society, 2,* 174–189.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review, 105,* 803–816. doi:10.1175/1520-0493(1977)105<0803:TVOCCA>2.0.CO;2

Nadav-Greenberg, L., & Joslyn, S. (2009). Uncertainty forecasts improve decision-making among nonexperts. *Journal of Cognitive Engineering and Decision Making, 2,* 24–47. doi:10.1518/155534308X284354

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.

Nightingale, S. D. (1987). Risk preference and laboratory use. *Medical Decision Making, 7,* 168–172. doi:10.1177/0272989X8700700307

Nohre, L. (2001). The effects of false alarms on driving behavior. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 62,* 2523.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230–253. doi:10.1518/001872097778543886

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 Forecasting, 21,* 116–122. doi:10.1175/WAF887.1

Seigrist, M., & Gutscher, H. (2008). Natural hazards and motivation for mitigation behavior: People cannot predict the affect evoked by a severe flood. *Risk Analysis, 28,* 771–778. doi:10.1111/j.1539-6924.2008.01049.x

Sloughter, J. M., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review, 135,* 3209–3220. doi:10.1175/MWR3441.1

Slovic, P. (1999). Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. *Risk Analysis, 19,* 689–701. doi:10.1111/j.1539-6924.1999.tb00439.x

Slovic, P., Monahan, J., & MacGregor, D. G. (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,* 271–296. doi:10.1023/A:1005595519944

Smith, S. K., & McCarty, C. (2009). Fleeing the storm(s): An examination of evacuation behavior during Florida's 2004 hurricane season. *Demography, 46,* 127–145. doi:10.1353/dem.0.0048

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

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representations of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323. doi:10.1007/BF00122574

Weber, E. U. (2006). Evidence-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change, 77,* 103–120. doi:10.1007/s10584-006-9060-3

(*Appendix follows*)

Appendix

Instructions to Participants in the Decision Aid Formats, Experiment 2

Decision Aid
"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."
Decision Aid Calculation *same as decision aid, plus the following:*
"Here is how DSA decides when to tell you to salt:
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).
● Salting costs $1,000.
● Not salting costs nothing, unless a freezing temperature is observed, which results in a penalty of $6,000.
● Therefore, the cost of not salting is actually $6,000 x probability of freezing.
● Weighing the cost of salting against the cost of not salting, there is a break-even point in probability of freezing of 16.67% ($1,000 / $6,000 = .1667, or 16.67%).
● 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.
The DSA accounts for probability of freezing and tells you:
Salt if probability of freezing ≥ 17%
Do not salt if probability of freezing < 17%"
Decision Aid Explanation *same as decision aid, plus the following:*
"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.
The advice will be given in each trial."

Received November 2, 2010
Revision received July 1, 2011
Accepted July 18, 2011 ■