DOI: 10.1111/risa.12336

The Cry Wolf Effect and Weather-Related Decision Making

Jared LeClerc and Susan Joslyn*

Despite improvements in forecasting extreme weather events, noncompliance with weather warnings among the public remains a problem. Although there are likely many reasons for noncompliance with weather warnings, one important factor might be people's past experiences with false alarms. The research presented here explores the role of false alarms in weather-related decision making. Over a series of trials, participants used an overnight low temperature forecast and advice from a decision aid to decide whether to apply salt treatment to a town's roads to prevent icy conditions or take the risk of withholding treatment, which resulted in a large penalty when freezing temperatures occurred. The decision aid gave treatment recommendations, some of which were false alarms, i.e., treatment was recommended but observed temperatures were above freezing. The rate at which the advice resulted in false alarms was manipulated between groups. Results suggest that very high and very low false alarm rates led to inferior decision making, but that lowering the false alarm rate slightly did not significantly affect compliance or decision quality. However, adding a probabilistic uncertainty estimate in the forecasts improved both compliance and decision quality. These findings carry implications about how weather warnings should be communicated to the public.

KEY WORDS: Cognitive psychology; decision making; false alarm effect; risk communication

1. BACKGROUND

With advances in atmospheric science, weather warnings are becoming increasingly timely and accurate.(1,2) Despite these facts, injury and loss of life still occur. Particularly devastating were the tornado outbreaks of 2011. Excellent National Weather Service forecasts with substantial lead times did not prevent massive loss of life.(3) This and other equally dramatic examples have led to a growing consensus that a substantial proportion of the problem resides in psychological factors governing public response to warning forecasts. Indeed, there is now considerable evidence for poor response to weather warnings of all kinds. In a recent study analyzing the tornado seasons of 2009–2011, Nagele and Trainor(4) found that the likelihood of seeking shelter was no greater for those within the warning polygon than for those living outside of it in the same county (about 40%). The compliance rate for hurricane warnings can also be unacceptably low. Interviews of residents under mandatory evacuation for Andrew and Hugo, both Category 4 hurricanes, revealed that only 42% evacuated their homes.(5) For Hurricane Floyd, the evacuation rate among those sampled was 64%.(6) 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.(7) Similar reluctance to take precautionary action has been observed for flood warnings.(8–11)

Although multiple factors are likely to blame,(12–14) this troubling noncompliance may arise at least in part from a general psychological tendency toward risk seeking in situations that involve a cost. Residents often regard precautionary action, like evacuation, as costly, citing reasons

University of Washington, Dept. of Psychology, Box 351525, Seattle, WA 98195, USA; (E-mail: jleclerc@uw.edu.)
*Address correspondence to Susan Joslyn, University of Washington, Dept. of Psychology, Box 351525, Seattle, WA 98195, USA; susanj@uw.edu.

such as travel expenses, dangers faced on the highway, inconvenience, and loss of property due to looting.[15–17] Moreover, decisions must be made early when the probability of adverse weather for any given location is low, often below 50%. Evidence suggests that in a variety of situations with these characteristics, some people, including both experts and nonexperts, assume more risk than is economically rational.[18–20] Thus, reluctance to evacuate may be due at least in part to a more general psychological tendency toward risk seeking in situations where avoiding risk involves a cost.[21]

However, this may not be the whole story when it comes to public response to weather warnings. Reluctance to take action may also be due in part to lack of trust in the warning. Survey evidence suggests that mistrust plays a major role in public attitudes toward scientific risk assessment in general[22] and warning forecasts in particular.[23,24] Warning forecasts may be particularly suspect because of their high false alarm (FA) rate, potentially leading to what is known as the "cry wolf" effect,[25] in which people hesitate to respond to subsequent warnings due to prior experience with FAs. Indeed, because of the high costs associated with a miss, the predominant error for warning forecasts is a FA.[26,27] If the cry wolf effect exists in weather warning situations, then a potential solution might be to increase the threshold for warnings slightly, which would reduce FAs. Fewer FAs might increase compliance to warnings overall,[28] offsetting any costs due to the slight increase in misses. This solution was tested in the research reported here. An alternative solution would be to change how weather warnings are communicated, a solution also explored in the present research.

Although there has been much discussion of the impact of FAs on trust in weather warnings, the psychological effects of prior experience with FAs remain unclear.[26] Among studies that investigated the impact of FAs in natural settings, there is some evidence suggesting that people are fairly tolerant of such errors,[13] especially if the cause of the FA is understood.[29] Other evidence suggests that people prefer to make their own decisions rather than relying on a warning system after a FA, although willingness to evacuate was not affected.[23] However, some evidence suggests a classic cry wolf effect, such as reluctance to heed future alarms following an unrealized earthquake prediction.[30] Anecdotal evidence also suggests that FAs are influential. A prime example is the rare snowstorm in Atlanta, GA, in January 2014. Despite clear and timely warnings given by the National Weather Service,[31] officials decided against closing schools and government agencies in advance,[32] resulting in approximately 1 million motorists attempting to leave the city at the same time on icy roadways and an 18-hour traffic jam. The potential for a FA may well have been an important factor among several others in this complicated and quickly evolving emergency situation. Indeed, in a press conference that evening, Governor Nathan Deal justified the decisions against advance closings by citing concern about FAs: "We don't want to be accused of crying wolf."[33] On the whole, however, direct evidence for the cry wolf effect is mixed.

The inconsistency among the studies reviewed above may be due to multiple uncontrolled variables that affect survey respondents. For instance, the exact nature of respondents' prior exposure to FAs, including the number of prior FAs, is often unknown. There is some evidence that the cry wolf effect may not be apparent after a single FA, arising only after several FA experiences.[34] Furthermore, the degree to which a warning results in a FA can vary.[26] In some cases, a less severe weather event may have been experienced while in others the respondent may have experienced no severe weather at all. In still other cases the severe event may have occurred in a neighboring location.

Thus, the impact of FAs on willingness to take precautions may be a question that is best answered in a controlled laboratory experiment. Exerting experimental control allows researchers to systematically manipulate the rate and degree of prior exposure to FAs to observe the effects. Indeed, there is some important experimental work that does just that. In a series of experiments,[25] participants faced the threat of a painful electric shock. Following the cancellation of the threat in an initial trial, participants were significantly less worried about being shocked in a subsequent trial, as measured by heart rate, subjective ratings of tension, and the credibility of subsequent threats, suggesting a clear FA effect. The effect was especially pronounced among participants who experienced the threat cancellation at a later stage of the initial trial and among participants who were told that there was a very high chance of the threat materializing. Importantly, when given the opportunity to pay (by way of reduced cash earnings) to reduce the intensity of the anticipated shocks, participants elected to take such protective action significantly less after experiencing FAs. While this is clear evidence for a FA effect, there are a number of differences between the decisions in these studies and those resulting from weather warnings that prevent a direct generalization. The threat of electric shock, for instance, is an unfamiliar and artificial situation with which participants had no prior experience and over which they had little or no control. Weather, on the other hand, is something with which people have vast prior experience, and specific courses of action are available to most people facing severe weather. Moreover, the primary dependent variables in these studies were emotional and physiological responses, whereas the critical response in real-life situations is often a binary go-no go decision.

There are also more ecologically valid experimental studies exploring FAs. These involve automated warning systems, many of them driving simulations. They also provide evidence for fairly strong FA effects in terms of failing to respond or responding less frequently,[35–37] responding more slowly,[38,39] using less efficient problem-solving strategies, and less frequent monitoring,[40] as well as reduced trust.[35,41] However, these are also not directly generalizable to weather warning situations due to a number of key differences. In automated systems tasks, the alarms are signals arising from and monitoring mechanical processes and precautionary action is often as simple as reducing one's speed or pressing a button. The consequences of a single miss are usually minor. Moreover, they are often tested in a dual task paradigm in which responding to the alarm is paired with another independent but sometimes related task. Weather warnings, on the other hand, are predictions about specific future events. Deciding whether to take precautionary action is generally the primary or only task during the relevant time frame, the process is more deliberate and complex, and the consequences are more costly to the decisionmaker. Thus, people may react quite differently to FAs in weather-related decision tasks than they do in the experimental studies reported to date.

For that reason, it is important to test this question in a controlled experimental paradigm using realistic weather scenarios and to test the impact of FA rate on the weather-related decisions of individual decisionmakers. That was one of the major goals of the experiment reported here. We sought to determine whether an *increase* in FAs would decrease participants' willingness to take precautionary action. In other words, is there a cry wolf effect for weather-related decision making? The second goal was to assess the impact of *reducing* FAs on compliance. Fewer FAs may preserve trust in the forecast and increase compliance with warnings. If so, minimizing FAs might help to increase compliance in actual weather warnings.

However, there is another factor that may impact compliance with weather warnings: the forecast information itself. It is possible that including an uncertainty estimate in the forecast will preserve trust despite a high rate of FAs because the uncertainty is acknowledged initially. This may in turn lead to better compliance. At present, despite advancements in atmospheric science leading to reliable uncertainty estimates for many weather parameters,[42–44] numerical probabilities are not often included in severe weather forecasts. Recent behavioral research[20] suggests that probabilistic forecasts for nighttime low temperature lead to better decisions than do deterministic forecasts because they allow people to better differentiate between situations that do and do not require precautionary action. This advantage increased when the error in the single-value forecast increased, suggesting that acknowledging uncertainty in the forecast combats the negative impact of forecast error. Indeed, participants expressed greater trust in the probabilistic forecasts. Thus uncertainty information may preserve trust in the face of FAs as well.

In addition, uncertainty information may provide users with a valid understanding of the risks they are facing. People realize that weather forecasts involve uncertainty even when they are presented as deterministic and attempt to estimate the uncertainty themselves.[27,45,46] Indeed, everyday users anticipate a wide range of values for deterministic forecasts and a high FA rate for weather warnings, and they regard extreme forecasts as exaggerations.[27,47] As a result of these intuitions, people may actually underestimate the risk in some situations if a valid uncertainty estimate is not provided. Furthermore, they may regard warnings that exclude uncertainty estimates as incomplete and untrustworthy. This combination of factors may play a role in noncompliance with weather warnings. If so, then forecasts that include an uncertainty estimate may improve decisions, preserve trust, and increase compliance. Testing this hypothesis was the third goal of the experiment reported here.

The experiment described below was designed to answer these three important questions by systematically manipulating the FA rate to determine (1) whether an increase in FAs decreases willingness to take precautionary action and trust, a cry wolf effect, (2) whether a decrease in FAs increases willingness to take precautionary action and trust, and (3) whether including an uncertainty estimate with the forecast attenuates any negative effects of FAs.

The experiment reported below employs a task in which participants decided whether or not to apply salt brine to a town's roads to prevent icing.[20,48] The decision, a simplified version of the real-world task, was based on an overnight low temperature forecast and the recommendations of a fictional automated decision aid. The FA rate was manipulated by varying the probability of freezing at which the decision aid advised applying salt treatment.

2. METHOD

2.1. Participants

A total of 388 University of Washington psychology students (54.9% female) participated for course credit and the chance to earn prize money. The mean age was 19.3 years (range 18–36 years). The majority of participants (80.4%) reported that they usually thought about temperature in degrees Fahrenheit.

2.2. Apparatus

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

2.3. Procedure

Participants were tested in small groups (1–12) in a computer lab. After they gave informed consent and provided demographic information, participants read a set of instructions at the same time that the experimenter read them aloud. The instructions 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. Applying salt brine cost $1,000 per day. There was a penalty of $5,000 for failing to apply salt brine when a freezing temperature was observed. There were 60 trials representing the days in two winter months. Participants received a virtual monthly budget of $35,000 and were instructed to attempt to maximize profits by minimizing salting expenses and avoiding penalties.

Salt brine had to be applied before freezing temperatures were reached. On each trial, representing one day, a forecast for the nighttime low temperature appeared on the screen. In addition, some participants were given the advice of a decision support aid, described to them as an advanced computer modeling system that would provide salting recommendations by combining information about the forecast, the uncertainty, the cost of salting, and the penalty for not salting. The decision support aid recommended applying salt on trials in which the probability of a freezing temperature was above a certain threshold ("Applying salt brine is recommended under these circumstances"), and it recommended against applying salt on trials in which the probability of freezing was below a certain threshold ("No action is recommended under these circumstances"), discussed below. The forecast and advice remained on the screen until the end of the trial, which involved three steps. First, participants indicated their trust in the forecast on a six-point drop-down menu, ranging from "not at all" to "completely." Next, they made their decision by clicking on one of two buttons marked "Salt" or "Not salt." Finally, participants indicated what they thought the nighttime low temperature would be by entering a numeric value in a text box. Immediately afterward, the observed nighttime low temperature and any balance 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 indicated their overall trust in the forecasts. Participants clicked "Next" to continue on to the next month's trials, and $35,000 was added to their balance. At the end of the final "month," participants received a cash payment commensurate with their ending balance. They received $3 for the ending balance amount that would result from following the advice on every trial (although it was described to participants only in terms of dollar amounts) and an additional dollar for each additional $5,000 above that balance (see Table I). Experimental sessions lasted approximately 45 minutes.

2.4. Design

A 3×4 incomplete factorial between-participants design was used. The first independent variable was forecast format. Participants were randomly assigned to one of three formats, all of which included the deterministic forecast. The control condition included

Table I. False Alarm Rates and Final Values Associated with Perfect Compliance at the Four False Alarm Levels

False Alarm (FA) Level	Number of Trials Salting Advised	Number of False Alarm Trials	False Alarm Rate	Perfect Compliance Expected Value (Mean)	Perfect Compliance Ending Balance
Higher FA	56	38	68%	−$962.50	$14,000
Unadjusted	45	29	64%	−$920.00	$15,000
Lower FA	30	18	60%	−$962.50	$10,000
Lowest FA	15	8	53%	−$1137.50	$0

only the deterministic forecast. The two experimental conditions also included the advice of the decision support aid. One of them included the probability of freezing (e.g., "there is a 22% chance that the temperature will be less than or equal to 32 °F") as well.

The second independent variable, manipulated only in the experimental conditions, was the FA level. In this task, FAs were trials in which salting was recommended by the decision support aid but the observed temperature was above 32 ºF. The FA level was manipulated by varying the probability of freezing at which the decision support aid recommended salting (see Table II). There were four levels with approximately equal numbers of participants in each FA level condition. One group of participants was advised to salt at and above the economically rational probability threshold (20%). The economically optimal threshold was determined by the cost-loss ratio, the cost of salting, and the expected value of the penalty for not salting.[49] At 20% chance of freezing, the probability-weighted penalty was equal to the cost of salting: 0.20 (probability) $\times$ \$5,000 (penalty) = \$1,000 (cost of salting). We will refer to this as the Unadjusted FA group. Another group of participants was advised to salt whenever the probability of freezing was 30% or greater, referred to as the Lower FA group because raising the threshold resulted in fewer FAs than using the economically optimal threshold. Another group of participants was advised to salt whenever the probability of freezing was 40% or greater, referred to as the Lowest FA group because it resulted in the fewest FAs. The final group of participants was advised to salt whenever the probability of freezing was 10% or greater. This was referred to as the Higher FA group because lowering the threshold resulted in more FAs than using the economically optimal threshold. The virtual cost-loss ratio remained the same in all conditions. However, in an effort to be fair to participants, the payout schedule in actual cash was based on reaching the ending balance that would be achieved by following the advice at each level (see Table I). Thus the payment schedule was commensurate with performance, taking into account differences in the optimal balance that could be achieved by following the recommendations in each condition.

2.5. Stimuli

Participants in all seven conditions received the same deterministic nighttime low temperature forecasts and observed temperatures in the same order. The ranges of temperature, probabilities of freezing (PoF), and forecast error 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 observed temperatures ranged from 26 °F to 41 °F ($M = 34$ °F). The sequence of observed temperatures followed a natural pattern with no difference from one night to the next exceeding 16°F. The mean standard forecast error was 3.17 ºF. Half of all observed temperatures were above their respective deterministic temperature forecasts and half were below. The PoF ranged from 8% to 51% (M = 29.0%). The probabilistic forecasts were reliable. The 60 forecasts were divided into four range categories (8–19%, 20–29%, 30–39%, 40–51%) and the percentage of observed temperatures 32 °F or less in each category remained within that probability range. For example, in the 20–29% range, temperatures 32 °F or less were observed on 4 of 15 (26.7%) days.

3. RESULTS

Before data analysis, we omitted participants whose temperature estimates or salting strategy suggested that they were not paying attention to the forecasts or not taking the task seriously. We subtracted the forecast value from participants' estimate

Table II. Decision Support Aid Decision Recommendation Thresholds Associated with Different False Alarm Levels

False Alarm (FA) Level	Freeze Probability Threshold Used by Decision Support Aid
Higher FA	Recommended applying salt if freeze probability ≥10%; recommended against applying salt if freeze probability <10%
Unadjusted	Recommended applying salt if freeze probability ≥20%; recommended against applying salt if freeze probability <20%
Lower FA	Recommended applying salt if freeze probability ≥30%; recommended against applying salt if freeze probability <30%
Lowest FA	Recommended applying salt if freeze probability ≥40%; recommended against applying salt if freeze probability <40%

on each trial. Participants whose mean temperature estimate differences were two or more standard deviations from the mean difference were removed. Additionally, participants who salted on every trial were removed from analysis. Based on these criteria, a total of 34 participants were removed, leaving 354 participants (56.2% female) for subsequent analysis.

The primary question for this research was whether an increase in FAs led to a decrease in compliance with the advice, a cry wolf effect. Compliance was calculated by creating a ratio comparing participants' decisions that coincided with the advice in the experimental conditions to advice-matching decisions in the control condition in which no advice was offered. Advice-matching decisions in the control condition were slightly different for each FA level condition because the advice in each condition recommended salting on slightly different trials depending on the probability of freezing threshold for that condition. A ratio greater than 1.00 indicated that participants made more advice-matching decisions than did participants without advice. A ratio that was not significantly different than 1.00 indicated that participants were uninfluenced by the advice. As Table III shows, on average ratios were greater than 1.00 in almost all experimental conditions. However, six one-sample t-tests (Bonferroni corrected p value of 0.0083) revealed that ratios were only significantly greater than 1.00 in the Unadjusted FA condition and the Lower FA condition that included the probability of freezing. This suggests that when the threshold was at the normative level or slightly above (when the probability of freezing was included), participants tended to comply with the advice. However, when the threshold was lower, yielding the most FAs, or much higher, yielding more misses, participants did not tend to heed the advice.

An ANOVA conducted on compliance ratios, with forecast format (Advice, Advice + Freeze Probability) and FA level (Higher, Unadjusted, Lower, Lowest) as the independent variables, revealed a significant cry wolf effect. There was a main effect of FA level, $F(3, 315) = 12.52$, $p < 0.01$. Participants complied with the advice significantly less often at the Higher FA level than at the Unadjusted FA level ($p = 0.01$, Cohen's $d = 0.50$) or Lower FA level ($p < 0.01$, Cohen's $d = 0.55$). Participants also complied with the advice significantly less often at the Lowest FA level than at the Unadjusted ($p = 0.01$, Cohen's $d = 0.78$) or Lower FA levels ($p = < 0.01$, Cohen's $d = 0.83$; see Table III). There was a significant main effect for forecast format, $F(1, 315) = 7.54$, $p < 0.01$. Those with probability forecasts followed the advice significantly more often ($M = 1.13$) than did those without them ($M = 1.06$; Cohen's $d = 0.32$).

It is important to note that although there was a main effect for FA level, according to Tukey's *post hoc* analysis, reducing FAs by raising the threshold slightly to 30% in the Lower FA condition did not improve compliance significantly compared to the Unadjusted condition ($p = 0.98$, Cohen's $d = 0.08$). Moreover, as we noted above, reducing FAs further in the Lowest condition actually *reduced* compliance significantly. Taken together, these results suggest that although there is a cry wolf effect, lowering the FA rate below the economically optimal threshold does not help and indeed eventually, at high enough levels, clearly hurts.

These results suggest further that a better way to improve compliance is to add an uncertainty estimate. Indeed, compliance at the Unadjusted FA level (64%) among participants who were given the probability of freezing was significantly better than compliance among those without the probability freezing at the Lower FA level (60%), $t(74) = 3.17$, $p < 0.01$, Cohen's $d = 0.74$.

Including the probability of freezing also led to better decisions. All decisions were assigned an

Table III. Mean and *SD* of Compliance Ratios, and Percent of Cases Greater than One (Indicating Being Influenced by Advice) by Format and False Alarm Level

Forecast Format	Higher FA	Unadjusted	Lower FA	Lowest FA	Overall
Advice	$M = 1.049$	$M = 1.108$	$M = 1.097$	$M = 0.988$	$M = 1.057$
	$SD = 0.23$	$SD = 0.17$	$SD = 0.15$	$SD = 0.26$	$SD = 0.21$
	$N = 38$	$N = 37$	$N = 36$	$N = 43$	$N = 154$
	$\% > 1 = 58\%$	$\% > 1 = 68\%$	$\% > 1 = 83\%$	$\% > 1 = 51\%$	$\% > 1 = 64\%$
Advice + FreezeProb	$M = 1.060$	$M = 1.201$*	$M = 1.231$*	$M = 1.004$	$M = 1.125$
	$SD = 0.26$	$SD = 0.13$	$SD = 0.16$	$SD = 0.23$	$SD = 0.22$
	$N = 37$	$N = 40$	$N = 42$	$N = 43$	$N = 162$
	$\% > 1 = 59\%$	$\% > 1 = 90\%$	$\% > 1 = 88\%$	$\% > 1 = 53\%$	$\% > 1 = 73\%$
Overall	$M = 1.055$	$M = 1.156$	$M = 1.169$	$M = 0.996$	$M = 1.092$
	$SD = 0.24$	$SD = 0.16$	$SD = 0.17$	$SD = 0.24$	$SD = 0.22$
	$N = 75$	$N = 77$	$N = 78$	$N = 86$	$N = 316$
	$\% > 1 = 59\%$	$\% > 1 = 79\%$	$\% > 1 = 86\%$	$\% > 1 = 52\%$	$\% > 1 = 69\%$

Table IV. Mean and *SD* Expected Loss by Format and False Alarm Level (Control Condition $M = \$1083.51$, $SD = 80.25$, $N = 38$)

Forecast Format	Higher FA	Unadjusted	Lower FA	Lowest FA	Overall
Advice	$M = \$1084.01$	$M = \$1049.55$	$M = \$1033.36$	$M = \$1042.07$	$M = \$1052.18$
	$SD = 83.73$	$SD = 81.97$	$SD = 64.47$	$SD = 52.89$	$SD = 73.22$
	$N = 38$	$N = 37$	$N = 36$	$N = 43$	$N = 154$
Advice + Prob	$M = \$1027.79$	$M = \$991.00$	$M = \$994.42$	$M = \$997.38$	$M = \$1001.99$
	$SD = 76.68$	$SD = 45.17$	$SD = 57.03$	$SD = 42.54$	$SD = 57.56$
	$N = 37$	$N = 40$	$N = 42$	$N = 33$	$N = 162$
Overall	$M = \$1056.28$	$M = \$1019.13$	$M = \$1012.39$	$M = \$1019.73$	$M = \$1026.45$
	$SD = 84.66$	$SD = 71.39$	$SD = 63.27$	$SD = 52.74$	$SD = 70.20$
	$N = 75$	$N = 77$	$N = 78$	$N = 86$	$N = 316$

expected loss. This was done by multiplying the penalty amount ($5,000) by the probability of freezing (the chance that penalty would be incurred) for each trial on which participants decided not to salt. On trials on which participants decided to salt, they were assigned the cost of salting ($1,000). Thus, *less* expected loss indicates *better* decision making. The mean expected loss for each condition is shown in Table IV.

An ANOVA on expected loss with both forecast format (Advice, Advice + Freeze Probability) and FA level (Higher, Unadjusted, Lower, Lowest) as the independent variables revealed a main effect for FA level, $F(3, 315) = 6.82$, $p < 0.01$, suggesting that decision quality declined as FAs increased. Tukey's *post hoc* tests revealed that those in the Higher FA condition did significantly worse than did those in the Unadjusted, Lower, and Lowest FA conditions, all $p < 0.01$, Cohen's $d = 0.47$, 0.59, and 0.52, respectively. No other differences were significant. This suggests that the primary impact of FA level was to reduce decision quality when FAs were increased. However, reducing FAs below the normative level did not lead to better decisions (lower expected loss).

Importantly, there was also a main effect for forecast format, $F(1, 315) = 47.10$, $p < 0.01$, indicating that decision quality was better when the advice was accompanied by a probability estimate. Indeed, adding the probability estimate to the Unadjusted FA condition led to lower expected loss than did lowering the threshold and decreasing FAs to 60%, $t(74) = 3.34$, $p < 0.01$, Cohen's $d = 0.76$.

Finally, we examined trust ratings. Recall the trust ratings were taken at the end of each month and on each day (i.e., each trial). In both the monthly and daily trust ratings, the forecast including the probability of freezing received the highest ratings. However, the monthly and daily trust ratings were slightly different with respect to FA level and compliance. The monthly trust ratings were clearly influenced by FA level: trust was higher when there were fewer FAs. An ANOVA on mean end-of-month ratings (see Table V) with forecast format

Table V. End-of-Month Trust Ratings (1-Not at All, 2-a Little, 3-Somewhat, 4-Quite a Bit, 5-Very Much, 6-Completely) by Format and False Alarm Level

Forecast Format	Higher FA	Unadjusted	Lower FA	Lowest FA	Overall
Advice	$M = 2.29$	$M = 2.42$	$M = 2.69$	$M = 2.57$	$M = 2.49$
	$SD = 0.87$	$SD = 1.07$	$SD = 0.80$	$SD = 0.97$	$SD = 0.94$
	$N = 38$	$N = 37$	$N = 36$	$N = 43$	$N = 154$
Advice + FreezeProb	$M = 2.38$	$M = 2.83$	$M = 2.77$	$M = 2.94$	$M = 2.74$
	$SD = 0.89$	$SD = 1.01$	SD = 0.99	$SD = 0.94$	$SD = 0.97$
	$N = 37$	$N = 40$	$N = 42$	$N = 43$	$N = 162$
Overall	$M = 2.33$	$M = 2.63$	$M = 2.74$	$M = 2.76$	$M = 2.62$
	$SD = 0.88$	$SD = 1.05$	$SD = 0.90$	$SD = 0.97$	$SD = 0.96$
	$N = 75$	$N = 77$	$N = 78$	$N = 86$	$N = 316$

(Advice, Advice + Freeze Probability) and FA level (Higher, Unadjusted, Lower, Lowest) as the independent variables revealed a significant main effect for FA level, $F(3, 315) = 3.24$, $p = 0.02$. Tukey's *post hoc* tests revealed that participants with the Higher FA level trusted the forecast significantly less than did those experiencing Lower ($p = 0.04$, Cohen's $d = 0.46$) and Lowest ($p = 0.03$, Cohen's $d = 0.46$) FA levels. There was also a significant main effect for forecast format, $F(1, 315) = 4.91$, $p = 0.03$, Cohen's $d = 0.26$, suggesting participants had greater trust when the probability of freezing was included. Interestingly, trust was not significantly correlated with compliance ratio in the Advice condition (r=0.13, p=0.16), although it was correlated in the Advice + Freeze Probability condition ($r = 0.22$, $p = 0.02$).

There was a closer relationship to compliance detected in the trial-by-trial trust ratings (see Table VI for means). Trust ratings were divided into two groups: (1) ratings made on trials during which participants complied with the advice ($M = 3.01$, $SD = 0.06$) and (2) ratings made on trials during which participant did not comply ($M = 2.58$, $SD = 0.05$). Then we conducted a mixed-model ANOVA on trust ratings. The within-participants factor was compliance versus noncompliance trials and the between-participants factors were forecast format (Advice, Advice + Freeze Probability) and FA level (Higher, Unadjusted, Lower, Lowest). Trust was significantly higher on compliance trials, $F(1, 314) = 172.65$, $p < 0.001$, and significantly higher in the condition in which the probability of freezing was included in the forecast ($M = 2.96$, $SD = 0.07$) compared to when it was not included ($M = 2.71$, $SD = 0.07$), ($F(1, 314) = 173.68$, $p < 0.001$). There was also a significant interaction between compliance and forecast format,

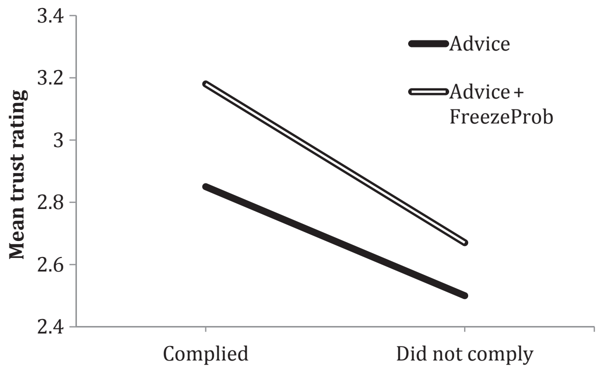

Fig. 1. Mean trial-by-trial trust ratings by forecast format for trials on which participants complied or did not comply with the decision advice.

$F(1, 314) = 5.71$, $p < 0.017$. As can be seen in Fig. 1, the difference in trust ratings between the two formats was much greater in compliance trials than noncompliance trials. Moreover, although the main effect for FA level did not reach significance, there was a significant interaction between FA level and compliance, $F(1, 314) = 3.34$, $p < 0.02$. As shown in Fig. 2, on compliance trials, there was a greater difference in trust ratings between FA levels and trust followed the FA level in the expected inverse order from lowest trust in the Higher FA condition to highest trust in the Lowest FA condition. On noncompliance trials there was less difference in trust ratings by FA levels and the Lower and Unadjusted FA levels were rated the lowest.

Taken together, these data suggest that what inspired the greatest trust overall was the forecast that included the probability of freezing. FA level had its strongest impact on trust in the monthly ratings, suggesting that these evaluations involved reflecting

Table VI. Mean Trial-by-Trial Trust Ratings (1-Not at All, 2-a Little, 3-Somewhat, 4-Quite a Bit, 5-Very Much, 6-Completely) by Format and False Alarm Level

Forecast Format	Higher FA	Unadjusted	Lower FA	Lowest FA	Overall
Advice	$M = 2.61$	$M = 2.64$	$M = 2.76$	$M = 2.87$	$M = 2.72$
	$SD = 0.91$	$SD = 0.83$	$SD = 0.78$	$SD = 1.01$	$SD = 0.89$
	$N = 38$	$N = 37$	$N = 36$	$N = 43$	$N = 154$
Advice + FreezeProb	$M = 2.91$	$M = 2.98$	$M = 3.06$	$M = 3.13$	$M = 3.02$
	$SD = 0.83$	$SD = 0.95$	$SD = 1.01$	$SD = 0.93$	$SD = 0.93$
	$N = 37$	$N = 40$	$N = 42$	$N = 43$	$N = 162$
Overall	$M = 2.76$	$M = 2.82$	$M = 2.92$	$M = 3.00$	$M = 2.88$
	$SD = 0.88$	$SD = 0.91$	$SD = 78$	$SD = 0.97$	$SD = 0.92$
	$N = 75$	$N = 77$	$N = 0.92$	$N = 86$	$N = 316$

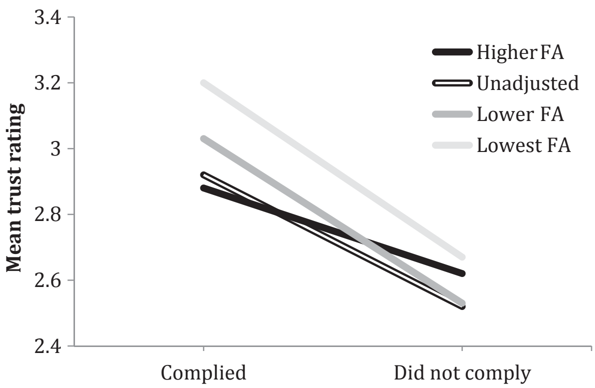

Fig. 2. Mean trial-by-trial trust ratings by false alarm level for trials on which participants complied or did not comply with the decision advice.

on the pattern of forecasts and outcomes over the entire month. However, daily trust ratings were more closely related to compliance, suggesting that trust evaluated at this level influences one's choices but that it is impacted primarily by the forecast information, i.e., higher ratings were awarded to the forecasts that included the probability of freezing.

4. CONCLUSION

This is the first evidence of which we are aware for a significant cry wolf effect in weather-related decision making using a controlled experimental approach. Participants were less likely to follow advice, trusted it less as reflected in the monthly ratings, and made economically inferior decisions when the advice led to more FAs (38 FAs out of 56 total salting recommendations, or 68%) compared to the level of FAs at the unadjusted normative threshold (29 FAs out of 45 total salting recommendations, or 64%). However, despite the fact that the manipulation was fairly strong, providing participants with nine more FA experiences over a short period of time, the effect size was moderate. In natural settings in which there is less prior experience with FAs spread over longer time periods, the effect may be much smaller and, as such, easily overpowered by other factors, leading to lack of clear evidence for the cry wolf effect in survey studies.

However, there was no evidence suggesting that lowering FAs (11 fewer FAs) increased compliance significantly. Moreover, lowering FAs further (21 fewer FAs) actually *reduced* compliance. This is probably because of the increase (5 more) in costly misses. The bottom line here is that FAs may indeed be a subtle contributing factor to noncompliance with weather warnings, but that lowering FAs below the economically optimal threshold may not help.

Interestingly, daily and monthly trust ratings seemed to be influenced by slightly different psychological factors. Monthly trust ratings were indeed higher for the decreased FA levels and, as such, appeared to be based on the pattern of forecasts and observations over the previous month. However, daily trust ratings were more closely related to compliance and influenced primarily by the information provided in the forecast. Participants had greater trust in forecasts that included the probability of freezing.

Indeed, perhaps the most important result of this experiment was the positive effect of the probabilistic forecast. The greatest increase in compliance, trust, and decision quality was achieved by adding a probability estimate to the forecast. Adding the probability of freezing led to greater compliance with the advice and greater increase in decision quality than did lowering the FA level. The implications for warning situations are important. In situations like this, when the cost of a miss is much greater than the cost of a FA, users will inevitably experience numerous FAs. This evidence suggests that adding an uncertainty estimate allows the forecast to retain its credibility despite this fact.

However, it is noteworthy that the type of uncertainty estimate may also be important. Although this variable was not tested in the experiment reported here, previous research suggests that benefits may depend on a match between uncertainty expression and the context. For instance, when precautionary action is required at fairly low event likelihoods (e.g., 20% chance), odds ratios, which describe the increase in odds in the present situation over climatological odds, may be preferable.[50]

In sum, the research reported here suggests that people can be fairly tolerant of FAs in weather-related decision tasks, as long as they are also provided an uncertainty forecast. Clearly, the decision processes underlying people's compliance with warnings is complicated, involving many factors, only some of which were tested here. Future research might explore additional factors associated with the real-world decision-making processes in weather emergency situations, such as time frame and whether or not decision outcomes affect the decision maker directly.

In sum, the implications of the research reported here for warning communication are important. They suggest that, in cases such as the 2014 Atlanta snowstorm, a numeric estimate conveying the high likelihood of the event may have convinced officials to close schools and government agencies earlier. Moreover, continued research on the cognitive factors that influence decision making in weather emergencies could play a critical role in increasing public compliance with warnings and reducing weather-related injury and death.

ACKNOWLEDGMENTS

This material is based on work supported by National Science Foundation Grant 0724721.

REFERENCES

1. Golden JH, Adams CR. The tornado problem: Forecast, warning, and response. Natural Hazards Review, 2000; 1(2):107–118.
2. National Hurricane Center. Annual NHC verification reports. Available at: http://www.nhc.noaa.gov/verification/verify3.shtml, Accessed April 2, 2014.
3. Paul BK, Stimers M. Exploring probable reasons for record fatalities: The case of 2011 Joplin, Missouri, tornado. Natural Hazards, 2012; 64(2):1511–1526.
4. Nagele DE, Trainor JE. Geographic specificity, tornadoes, and protective action. Weather, Climate, and Society, 2012; 4(2):145–155.
5. Riad JK, Norris FH, Ruback RB. Predicting evacuation in two major disasters: Risk perception, social influence, and access to resources. Journal of Applied Social Psychology, 1999; 29(5):918–934.
6. Dow K, Cutter SL. Public orders and personal opinions: Household strategies for hurricane risk assessment. Global Environmental Change Part B: Environmental Hazards, 2000; 2(4):143–155.
7. Gibbs LI, Holloway CF. Hurricane Sandy after action: Report and recommendations to Mayor Michael R. Bloomberg, 2013 Available at: http://www.nyc.gov/html/recovery/downloads/pdf/sandy_aar_5.2.13.pdf, Accessed April 2, 2014.
8. Perry RW. Population evacuation in volcanic eruptions, floods, and nuclear power plant accidents: Some elementary comparisons. Journal of Community Psychology, 1983; 11(1):36–47.
9. Gruntfest EC, Downing T, White G. Big Thompson flood exposes need for better flood reaction system to save lives. Civil Engineering, 1978; 48:72–74.
10. Parker DJ, Priest SJ, Tapsell SM. Understanding and enhancing the public's behavioural response to flood warning information. Meteorological Applications, 2009; 16(1):103–114.
11. Horney JA, MacDonald PD, Van Willigen M, Berke PR, Kaufman JS. Individual actual or perceived property flood risk: Did it predict evacuation from Hurricane Isabel in North Carolina, 2003? Risk Analysis, 2010; 30(3):501–511.
12. Whitehead JC, Edwards B, Van Willigen M, Maiolo JR, Wilson K, Smith KT. Heading for higher ground: Factors affecting real and hypothetical hurricane evacuation behavior. Global Environmental Change Part B: Environmental Hazards, 2000; 2(4):133–142.
13. Baker EJ. Hurricane evacuation behavior. International Journal of Mass Emergencies and Disasters, 1991; 9(2):287–310.
14. Dash N, Gladwin H. Evacuation decision making and behavioral responses: Individual and household. Natural Hazards Review, 2007; 8(3):69–77.
15. Blendon RJ. Hurricane Readiness in High-Risk Areas: May 27–June 23, 2008. Boston, MA: Harvard School of Public Health, 2008.
16. Cutter SL, Smith MM. Fleeing from the hurricane's wrath: Evacuation and the two Americas. Environment: Science and Policy for Sustainable Development, 2009; 51(2):26–36.
17. Smith SK, McCarty C. Fleeing the storm(s): An examination of evacuation behavior during Florida's 2004 hurricane season. Demography, 2009; 46(1):127–145.
18. Betts SC, Taran Z. The effects of age and reliability on price in the used car market: A test of prospect theory and multiple reference points. Academy of Marketing Studies, 2005; 10(1):11–15.
19. Haas ML. Prospect theory and the Cuban missile crisis. International Studies Quarterly, 2001; 45(2):241–270.
20. Joslyn SL, LeClerc JE. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. Journal of Experimental Psychology: Applied, 2012; 18(1):126–140.
21. Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. Econometrica, 1979; 263–291.
22. Slovic P. Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. Risk Analysis, 1999; 19(4):689–701.

23. Dow K, Cutter SL. Crying wolf: Repeat responses to hurricane evacuation orders. Coast Management, 1998; 26(4):237–252.
24. Morss RE, Hayden MH. Storm surge and ''certain death'': Interviews with Texas coastal residents following Hurricane Ike. Weather, Climate, and Society, 2010; 2(3):174–189.
25. Breznitz S. Cry Wolf: The Psychology of False Alarms, 1st ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1984.
26. Barnes LR, Gruntfest EC, Hayden MH, Schultz DM, Benight C. False alarms and close calls: A conceptual model of warning accuracy. Weather and Forecasting, 2007; 22(5):1140–1147.
27. Joslyn SL, Savelli S. Communicating forecast uncertainty: Public perception of weather forecast uncertainty. Meteorological Applications, 2010; 17(2):180–195.
28. Roulston MS, Smith LA. The boy who cried wolf revisited: The impact of false alarm intolerance on cost-loss scenarios. Weather and Forecasting, 2004; 19(2):391–397.
29. Sorensen JH. Hazard warning systems: Review of 20 years of progress. Natural Hazards Review, 2000; 1(2):119–125.
30. Atwood LE, Major AM. Exploring the "cry wolf" hypothesis. International Journal of Mass Emergencies and Disasters, 1998; 16(3):279–302.
31. National Weather Service. North and Central Georgia winter storm: January 28–29, 2014. Available at: http://www.srh.noaa.gov/ffc/?n=20140128winterstorm, Accessed March 15, 2014.
32. Sposito S. Winter storm 2014 timeline. Atlanta Journal-Constitution, 2014. Available at: http://www.ajc.com/news/news/winter-storm-2014-timeline/nc6k9/, Accessed March 15, 2014.
33. Severson K. Atlanta officials gamble on storm and lose, and others pay the price. New York Times, 2014. Available at: http://www.nytimes.com/2014/01/30/us/ice-storm-southern-united-states.html?_r=0, Accessed February 2, 2014.
34. Carsell K. Impacts of a false alarm: The January 29, 2000 Ventura, California experience (master's thesis). Colorado Springs, CO: University of Colorado at Colorado Springs, 2001.
35. Lees MN, Lee JD. The influence of distraction and driving context on driver response to imperfect collision warning systems. Ergonomics, 2007; 50(8):1264–1286.
36. Cotte N, Meyer J, Coughlin JF. Older and younger drivers' reliance on collision warning systems. Human Factors and Ergonomics Society Annual Meeting Proceedings, 2001; 45(4):277–280.
37. Bliss JP, Gilson RD, Deaton JE. Human probability matching behaviour in response to alarms of varying reliability. Ergonomics, 1995; 38(11):2300–2312.
38. Getty DJ, Swets JA, Pickett RM, Gonthier D. 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, 1995; 1(1):19–33.
39. Kerstholt JH. Decision making in a dynamic situation: The effect of false alarms and time pressure. Journal of Behavioral Decision Making, 1995; 8(3):181–200.
40. Kerstholt JH, Passenier PO. Fault management in supervisory control: The effect of false alarms and support. Ergonomics, 2000; 43(9):1371–1389.
41. Gupta N, Bisantz AM, Singh T. Investigation of factors affecting driver performance using adverse condition warning systems. Human Factors and Ergonomics Society Annual Meeting Proceedings, 2001; 45(23):1699–1703.
42. Gneiting T, Raftery AE. Weather forecasting with ensemble methods. Science, 2005; 310(5746):248–249.
43. Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 2005; 133(5):1155–1174.
44. Sloughter JM, Raftery AE, Gneiting T, Fraley C. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Monthly Weather Review, 2007; 135(9):3209–3220.
45. Morss RE, Demuth JL, Lazo JK. Communicating uncertainty in weather forecasts: A survey of the US public. Weather and Forecasting, 2008; 23(5):974–991.
46. Lazo JK, Morss RE, Demuth JL. 300 billion served: Sources, perceptions, uses, and values of weather forecasts. Bulletin of the American Meteorological Society, 2009; 90(6):785–798.
47. Savelli S, Joslyn SL. Boater safety: Communicating weather forecast information to high-stakes end users. Weather, Climate, and Society, 2012; 4(1):7–19.
48. Roulston MS, Bolton GE, Kleit AN, Sears-Collins AL. A laboratory study of the benefits of including uncertainty information in weather forecasts. Weather and Forecasting, 2006; 21(1):116–122.
49. Thompson JC. On the operational deficiencies in categorical weather forecasts. Bulletin of the American Meteorological Society, 1952; 33:223–226.
50. LeClerc JE, Joslyn SL. Odds ratio forecasts increase precautionary action in severe weather events. Weather, Climate, and Society, 2012; 4:263–270.