

# Communicating Uncertainty Information in a Dynamic Decision Environment<sup>✉</sup>

GALA GULACSIK,<sup>a</sup> SUSAN L. JOSLYN,<sup>a</sup> JOHN ROBINSON,<sup>b</sup> AND CHAO QIN<sup>a</sup>

<sup>a</sup> *Department of Psychology, University of Washington, Seattle, Washington*

<sup>b</sup> *Human Centered Design and Engineering, University of Washington, Seattle, Washington*

(Manuscript received 27 December 2021, in final form 2 August 2022)

**ABSTRACT:** The likelihood of threatening events is often simplified for members of the public and presented as risk categories such as the “watches” and “warnings” currently issued by National Weather Service in the United States. However, research (e.g., Joslyn and LeClerc) suggests that explicit numeric uncertainty information—for example, 30%—improves people’s understanding as well as their decisions. Whether this benefit extends to dynamic situations in which users must process multiple forecast updates is as yet unknown. It may be that other likelihood expressions, such as color coding, are required under those circumstances. The experimental study reported here compared the effect of the categorical expressions “watches” and “warnings” with both color-coded and numeric percent chance expressions of the likelihood of a tornado in a situation with multiple updates. Participants decided whether and when to take shelter to protect themselves from a tornado on each of 40 trials, each with seven updated tornado forecasts. Understanding, decision quality, and trust were highest in conditions that provided percent chance information. Color-coded likelihood information inspired the least trust and led to the greatest overestimation of likelihood and confusion with severity information of all expressions.

**KEYWORDS:** Social science; Tornadoes; Uncertainty; Forecasting; Decision-making

## 1. Introduction

Despite improved forecasts with increasing lead time, residents of tornado-prone areas in the United States continue to be injured or killed by tornadoes every year. There is a growing consensus that deadly outcomes such as this may be due at least in part to the influence of psychological and social factors on public response to warning forecasts (Lindell and Perry 2012; Lindell 2018). According to a study looking at tornado seasons 2008–10, the likelihood of taking shelter was no greater for those under a tornado warning than for those outside of the warning area living in the same county (Nagele and Trainor 2012). While there are many reasons for noncompliance, some beyond the control of residents, the effectiveness of the risk communication may well be a contributing factor. Here, we define risk as a function of the likelihood and value of a future event, usually involving loss (Eiser et al. 2012).

One way that the National Weather Service (NWS) currently communicates tornado risk to the public is by issuing either a tornado watch or a tornado warning. A “watch” means tornadoes are possible in and near the designated area. A “warning” means that a tornado is imminent or occurring and taking shelter is advised (NWS 2012).

### a. Forecast uncertainty

Although the current warning system fails to acknowledge it directly, meteorologists know that the probability of a

tornado varies geographically within the warned area and changes over time (Karstens et al. 2015). Nonetheless, at present this information is not made available to members of the public. Whether to communicate probabilistic information such as this is the subject of continued debate. Evidence suggests that people understand that all forecasts involve some level of uncertainty (Joslyn and Savelli 2010), due to their own prior experience with the reliability similar forecasts. For instance, in the tornado season of May 2011 to May 2014, more than one-half (58%) of the 132 tornado warnings were false alarms, that is, no tornado was observed anywhere within the warned area (NWS 2011). Moreover, research links false alarm ratios to subsequent tornado casualties (Simmons and Sutter 2009) suggesting a reduction in compliance as a result of increasing false alarms. Thus, residents may regard some warnings as “wrong” and ignore future forecasts (Ripberger et al. 2015). However, the relationship between false alarms and public perception of the validity such warnings is likely complex. For instance, perceived false alarm rate is not necessarily correlated with the actual false alarm rate and may depend on whether the false alarm is a “close call” (Barnes et al. 2007; Lim et al. 2019). Nonetheless, noncompliance with warnings may be due, at least in part, to lack of trust.<sup>1</sup>

If so, it may help to add an uncertainty estimate. Experimental evidence suggests that adding numeric uncertainty information (e.g., 30% chance) attenuates the reduction in trust due to increased error in forecasts (Joslyn and LeClerc 2012). In addition, there is evidence that adding explicit uncertainty information to the forecast preserves trust to a greater degree than does reducing false alarms (LeClerc and Joslyn 2015).

<sup>1</sup> In this context, we define trust as the degree to which the information appears reliable, adequate, or complete (Earle 2010).

<sup>✉</sup> Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WCAS-D-21-0186.s1>.

Corresponding author: Chao Qin, [robertqc@uw.edu](mailto:robertqc@uw.edu)

DOI: 10.1175/WCAS-D-21-0186.1

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy \(www.ametsoc.org/PUBSReuseLicenses\)](https://www.ametsoc.org/PUBSReuseLicenses).

Moreover, there is now strong evidence that the advantages for probabilistic forecasts go beyond trust, improving both understanding and decision quality as well (Joslyn et al. 2007, 2009; Joslyn and LeClerc 2012). Indeed, these advantages may be the reason that numeric uncertainty estimates increase trust: better understanding allows users to reach their decision goals.

#### b. Multiple forecast updates

However, there are some important differences between the experimental settings cited above and real-world weather warnings. In experimental settings participants are generally asked to make a decision on the basis of a single forecast. However, in natural settings forecasts often begin several days in advance of major events and are updated over time. In the case of a tornado, for instance, residents can receive multiple forecast updates in a single day as the situation continues to evolve. With each update the information may change, requiring residents to replace old with current information. Keeping track of rapidly changing information such as this can be challenging due to limitations in what is referred to as “working memory” capacity, our ability to maintain and update information in consciousness (Baddeley 2000). Because uncertainty information constitutes additional and somewhat more complex information than what is currently provided in most warnings, it may not be fully processed when rapid updates are provided. However, to our knowledge, no experimental research has yet tested this question. Therefore, the experiment reported here was designed to test the following hypotheses, in a simulated tornado warning situation with multiple forecast updates, to determine whether probabilistic forecasts remain beneficial when this complexity is added:

- 1) Participants will have a better understanding of the likelihood of a tornado when forecasts include explicit numeric likelihood expressions as compared with categorical expressions such as the conventional watch and warning format.
- 2) Participants will have greater trust in forecasts that include explicit numeric likelihood expressions as compared with categorical expressions.
- 3) Participants will make better decisions (for specific operationalizations<sup>2</sup> of this construct, see procedure section below) when forecasts include explicit numeric likelihood expressions as compared with categorical expressions.

#### c. Second-order uncertainty

The processing challenge may be compounded by the fact that assigning a single probability may not be possible for some weather events. In practice, a range of likelihoods such as 10%–20% chance of a tornado may be used. This is referred to as second-order uncertainty, an expression that indicates uncertainty about the uncertainty. Research suggests

that users can understand likelihood ranges and sometimes regard them as more credible than single probabilities (Dieckmann et al. 2010). However, likelihood ranges further increase the amount of information that must be processed and may be particularly challenging in a dynamic decision environment. Thus, another goal of the research reported here was to test whether probability ranges provide the same advantages as a single estimate.

#### d. Color coding

Because of the increased processing demands in dynamic decision environments like a tornado event, numeric expressions of likelihood such as probabilities or probability ranges may be too challenging. Therefore, it may be necessary to simplify likelihood expressions to allow for rapid and easy understanding. Many believe, for instance, that color coding, due to its high salience (Wogalter et al. 2002), may be appropriate. In fact, evidence suggests that a multihue color-coded expression of wildfire likelihood led to better decisions under time pressure, although simple text expressions were more advantageous otherwise (Cheong et al. 2016). Therefore, in the research reported here, in addition to testing numeric expressions of likelihood, we also tested color-coded likelihood expressions.

There are two main types of color coding, multihue as mentioned above, and a single hue, such as red, that varies in value. The research on these formats is mixed, but slightly favors multihue formats. Multihue schemes have been shown to be read faster and sometimes preferred by users although they are not significantly different in terms of reading accuracy than a red scale varying in value (Miran et al. 2017). However, some research suggests that users find a multihue scale *less* intuitively linked to uncertainty than a single hue with variation in value (MacEachren et al. 2012), although neither reading accuracy or speed of interpretation were tested in that research. Thus, it is not clear which color-coded format is superior overall.

Here we focus on a multihue color-coded scheme, as it has already been integrated into risk communication tools developed by the National Oceanic and Atmospheric Administration in the Forecasting a Continuum of Environmental Threats (FACETs) framework for tornado threats. Our main question was how accurately multihue uncertainty formats communicate likelihood. At present, the research addressing this issue is sparse. There is evidence that people prefer color coding, such as a “traffic light” (red, yellow, and green) configuration to represent uncertainty in such situations (Radford et al. 2013; Tak and Toet 2014). Color-coded likelihood has also been shown to correct the common but incorrect assumption that tornadoes are more likely to occur in the center of the warned area (Ash et al. 2014). Although it is difficult to discern whether this effect is due to the color coding per se, or to the contours that it creates. However, to date, as far as we are aware, there is no research that investigates whether users understand that color coding is intended to indicate likelihood alone when that is the case, or whether their interpretations match the precise level of likelihood intended by forecasters.

---

<sup>2</sup> By “operationalize” we mean to translate the theoretical construct into an overt measurable quantity.

We address both potential misunderstandings in the paragraphs below.

*e. Likelihood–severity confusion*

The fundamental question about people's understanding of color coding is what they construe the colors to represent. This is an issue due in part to the fact that color coding involves an extra step, reading a legend or explanation that translates the colors into an expression of something else, in this case likelihood. Whether or how carefully people engage in this step may be influenced by prior experience with color coding in other contexts that gives rise to expectations about what the colors mean. For instance, in the context of weather, color coding is often used to represent the severity of the event, that is, estimates such as wind speed or the amount of precipitation. In other cases, color coding is used to indicate risk, a combination of the likelihood and severity. Because of these precedents, users may have expectations about the meaning of the color coding that cause them to assume they already understand and misread or ignore the legend. As a result, users may misinterpret color-coded likelihood as an expression as severity or the extent of potential damage, or some combination of likelihood and severity, rather than as likelihood alone.

This inclination may be reinforced by the fact that even when color is not involved, there is evidence for a tendency to misinterpret graphically depicted likelihood information as some deterministic quantity such as wind speed or precipitation amount, referred to as a deterministic construal error (Joslyn and Savelli 2021). For example, users tend to misinterpret visualizations of percent chance of precipitation as duration or geographic extent of precipitation (Joslyn et al. 2009). In another study, bracket visualizations depicting the 80% predictive interval for nighttime low temperature [e.g., 35°–42°F (1.7°–5.6°C)] were misinterpreted as diurnal fluctuation. In other words, participants thought the endpoints of the range depicted two single-value forecasts, one for daytime high and the other for nighttime low (Savelli and Joslyn 2013). Similarly, many people think the cone of uncertainty, intended to show the possible hurricane path, depicts the extent of the wind field (Broad et al. 2007). Therefore, it may be that people have a general tendency to misinterpret likelihood expressions as some expression of severity if the expression permits it. This may have to do in part with cognitive load. A probabilistic forecast indicates that multiple outcomes are possible and thus, requires consideration of more information than a forecast describing the severity of a single outcome. It is possible that users tend to avoid the more difficult probabilistic interpretation and instead choose the easier severity interpretation, referred to as “attribute substitution” (Kahneman and Frederick 2002). One of the major goals of the research reported here is to determine the degree to which the likelihood–severity confusion arises from color-coded likelihood representations. Therefore, we add a prediction (in italics below) to hypothesis 1:

- 1) Participants will have a better understanding of the likelihood of a tornado when forecasts included explicit

numeric likelihood expressions as compared with categorical expressions such as the conventional watch and warning format or color coding. *Participants will tend to confuse likelihood and severity when likelihood is color-coded.*

*f. Understanding the level of likelihood*

The other issue about users understanding of color-coded likelihood, even when they understand that it is intended to portray likelihood, is whether color coding conveys the precise level of likelihood intended by forecasters. Here the research is sparse. Much of the work to date has been conducted on color-coded scales intended to indicate risk, a combination of likelihood and severity for which the dependent variable is the order in which participants rank colors to represent risk. The majority of the evidence suggests substantial variability in rank order, suggesting that there may be issues. Indeed, with the exception of red, often found to convey the notion of greatest risk (Borade et al. 2008; Hellier et al. 2010), there is little consensus on the rank order of colors to convey risk (Chapanis 1994; Wogalter et al. 1995; Griffith and Leonard 1997; Rashid and Wogalter 1997). In addition, there are cultural differences. Orange, rather than red, is considered of greatest hazard by Chinese participants (Lesch et al. 2009). Moreover, a study of the now retired Homeland Security Advisory System (HSAS), using color to indicate terrorist threat, showed that more than half of the participants (57.8%) ranked the colors from most threatening to least in an order that conflicted with that intended (Mayhorn et al. 2004). Therefore, color-coded risk may convey different levels of risk to different individual users as well as different levels of risk than what was intended.

It is important to note that the studies reviewed above tested the rank ordering of color-coded risk. It is possible that participants could accurately rank colors and still misunderstand the precise likelihood of the adverse event. In other words, orange may be ranked as the second highest risk and assigned 90% chance by the user when it was intended to represent 50% chance. Indeed, no studies of which we are aware have asked participants to assign likelihoods (e.g., “a 40% chance”) to each color category to determine how they compare with the intended values. Determining whether color coding conveys the likelihood intended is another of the major goals (see hypothesis 1, above) of the research reported here.

In sum, public trust in forecasts may be critical for appropriate and timely precautionary decisions in the face of severe weather. Moreover, trust may be maintained by providing understandable event likelihood information. Likelihood information may also improve decisions based on the forecast. At present however, whether these benefits will be seen in a dynamic decision environment and how best to convey likelihood information in that context remains an open question.

The experimental study reported here investigated these issues using a computerized decision task (based on Schwartz and Howell 1985). Participants decided whether to seek shelter based seven updated tornado forecasts presented

sequentially. Although more complex than the experimental scenarios tested previously, it was a vastly simplified version of the real-world task in which many other factors (e.g., geographic knowledge, environmental cues, prior experience), options (e.g., rearranging schedules, increasing monitoring), and information sources (e.g., social media) are considered at different stages (Lindell 2018). However, this simplification allowed for the experimental control necessary to infer direct causal relationships between the way in which the forecast information was expressed and differences in understanding, trust, and hypothetical shelter decisions. The conventional watch or warning format that served as our control condition was compared with numeric probabilistic forecasts (both single probabilities and ranges) and color-coded likelihood expressions.

## 2. Method

In the experimental study reported here, we compared the effects of various forecast formats (see stimuli section below) on participant understanding of forecast likelihood, trust in the forecast, and decision-making.

### a. Participants

Of the 489 University of Washington students who participated in the study for extra course credit, 57% were female. All participants were between 18 and 24 years of age, with an average age of 19 years. All were enrolled in a psychology course. Although racial information was not collected for this sample, the composition was likely similar to that of the student population in the year in which the experiment was conducted, 2019 (40% Caucasian, 25% Asian American, 8% Hispanic/Latino, 4% African American, 1% American Indian, and 0.9% Hawaiian/Pacific Islander; University of Washington 2019). Most participants were residents of Washington State, where tornados are rare. Evidence suggests that most University of Washington psychology students (80%) have no experience with tornados (C. Qin et al. 2021, unpublished manuscript).

### b. Procedure

Participants performed as computerized decision task (details described below) individually on desktop computers in a laboratory room that accommodated approximately 10 participants per session. The task was presented in an online html platform. The researcher read the instructions aloud (see the online supplemental material) as participants viewed the same instructions printed on the computer screens. The instructions provided background on the tornado hazard including how a tornado is formed as well as the potential damage to homes and occupants. To familiarize participants with the range of tornado hazards, instructions described the wind speeds for both weak (73–112  $\text{mi h}^{-1}$ ;  $1 \text{ mi h}^{-1} \approx 0.45 \text{ m s}^{-1}$ ) and strong (260+  $\text{mi h}^{-1}$ ) tornadoes. Then, participants were told that they would receive forecasts for 40 storms with the potential to produce tornadoes. Severity was held constant by informing participants that all storms would produce wind speeds of 90–112  $\text{mi h}^{-1}$ , consistent with weak tornadoes. Every storm moved west to east from the same distance, toward

TABLE 1. Cost of decisions (based on Schwartz and Howell 1985).

Decision	Cost
Wait	Forecast updates 1–3: no cost Forecast updates 4–7: 20 points per wait decision
Take shelter	Shelter cost = $300 + [3 \times (\text{forecast update})^2]$
Not take shelter	No cost, but 1500-point penalty if a tornado hits home

“home”—a house in which participants were to imagine they were located. Each storm constituted a single trial. One trial comprised seven successive forecast updates as the storm moved from west to east. Participants performed 10 practice trials followed by 40 data collection trials.

### 1) COST–LOSS STRUCTURE

To motivate participants to put forth their best effort, they were given a point balance (24 000 points) at the beginning of the simulation to spend on protective actions. The goal was to complete the task with as many points as possible. There were three decision options at each forecast update: wait, take shelter, and not take shelter (see Table 1). Participants could wait for more information at no cost on forecast updates 1–3. However, on forecast updates 4–7, there was a 20-point cost for every wait decision to reflect the increasing danger of the storm approaching home. Participants could choose to go to a nearby tornado shelter at a cost of 303 points at forecast update 1. This cost increased slightly (see Table 1) at each forecast update to reflect the increasing danger of being caught in a vulnerable position when a tornado strikes. All costs were deducted immediately from the onscreen point balance. Third, participants could choose for no cost to not take shelter. However, if a tornado hit home and the participant had chosen to not take shelter or wait as their final decision, a 1500-point penalty was immediately deducted from their balance. Notice that, in this task, as in actual weather hazard situations, the cost to protect oneself is much less than the loss that could result from inaction if a tornado hit the residents location.

### 2) STORM MOVEMENT

Participants saw different forecast updates depending on the path of the storm on a virtual grid of longitude and latitude shown in Fig. 1, although no geographic representation of storm movement was shown to participants. On each trial, the storm moved across the grid toward the participant’s home at the far eastern boundary. As the storm reached each new longitude, a forecast update was issued. At the start of every trial, the storm was located at latitude 4, longitude 1. From that point, storm movement was randomly generated in real-time but constrained to proceed from west to east and remain within the boundaries of the grid (see Fig. 1). When the storm was located in a cell between latitudes 2–6, there was a 0.3 probability of moving to the cell northeast or southeast and a 0.4 probability of moving laterally to a cell to the east. When the storm was located along on the top or bottom of the grid (latitudes 1 or 7) it advanced laterally to the east with

		Longitude							
		1	2	3	4	5	6	7	8
Latitude	1								
	2								
	3					X			
	4	X			X		X		Home
	5		X	X				X	
	6								X
	7								

FIG. 1. Example of storm movement across the grid. This graphic was not shown to participants.

a 0.7 probability and toward the center (i.e., southeast from latitude 1 or northeast from latitude 7) with a 0.3 probability. Thus, storm movement was more constrained in this experimental than in the real world, and for no storms was there reduced likelihood not related to the storm path (e.g., dissipation) as there might be in the real world. The probability of the storm being in each location on the grid is shown in Fig. 2.

### 3) TRIAL STRUCTURE

On each trial participants received a series of updated forecasts each time the storm advanced (seven per trial). At each update they answered two questions that reflected their understanding of the information provided. They rated the likelihood of a tornado hitting their home by clicking on a visual analogy scale (VAS), anchored on left end with “impossible” and the right end with “certain” (see Fig. 3). Participants also rated the damage they would expect if a tornado were to hit home on a similar VAS anchored on the left with “not severe” and the right with “very severe.” On the next screen, the same forecast was displayed, and participants indicated their shelter decision by clicking on one of three radio buttons labeled “wait,” “not-take-shelter,” or “take-shelter,” shown in that order. The cost for each choice was shown beside it (see Table 1). On the same screen, in order to gauge the impact of the information format on trust, participants rated how much they trusted the forecast by clicking on a VAS, anchored on the left with “not at all” and to the right with “completely.”<sup>3</sup> Then, an updated forecast was shown, and they answered the same set of questions with respect to the update. A decision to take shelter or not take shelter was a final decision for that storm (trial). However, in order to discourage rushing through trials by deciding early, when participants made a final decision prior to the seventh update, they saw the remaining forecast updates and answered the same questions about likelihood, severity, and trust although they were not allowed to change their decision. Note that

<sup>3</sup> Confidence was also measured, although no significant differences were detected and thus it will not be mentioned further.

		Longitude							
		1	2	3	4	5	6	7	8
Latitude	1	0	0	0	0.027	0.0513	0.0715	0.0875	0.1
	2	0	0	0.09	0.108	0.1188	0.1248	0.1291	0.1322
	3	0	0.3	0.24	0.225	0.2064	0.1923	0.1812	0.1725
	4	1	0.4	0.34	0.28	0.247	0.2226	0.2044	0.1905
	5	0	0.3	0.24	0.225	0.2064	0.1923	0.1812	0.1725
	6	0	0	0.09	0.108	0.1188	0.1248	0.1291	0.1322
	7	0	0	0	0.027	0.0513	0.0715	0.0875	0.1

FIG. 2. Probability of storm being in each location.

because there were seven updates, there were seven opportunities to make a final decision. Therefore, it was possible to evaluate both when the participant made a final decision (to which we refer as “timeliness”) as well as the decision itself. After all seven forecast updates, the outcome screen informed participants whether the tornado hit or missed the home, reminded them of their final choice and the cost, whether sheltering was “necessary” (if the tornado struck home), and whether they incurred or avoided a 1500-point penalty. Another trust rating was taken on the next screen, which also included the final score. Then, the next trial with the same sequence of screens was shown. Participants were told that each trial represented a storm independent from the others to discourage them from deducing trends in the weather conditions. After the last trial, the final screen showed the ending point balance, the number of times a tornado hit home, and a summary of the participant’s choices.

### 4) COMPENSATION

Participants were given course credit and rewarded \$1 for every 1500 points above 11 880 points remaining in their point balance at the end of 40 trials. This payout threshold was designed to further discourage the simplistic and unrealistic strategy of sheltering at the first update on all trials (24000 initial point balance – 12 120 spent to shelter = 11 880).

#### c. Stimuli

All forecasts were based on the probability of a tornado hitting home from the cell in which the storm was currently located (Fig. 4). The probabilities of a tornado hitting home in this simulation (to which we refer as “actual probability”) were realistic for a geographic region under tornado threat and ranged from 0 to 0.40.<sup>4</sup>

Participants saw one of five formats, a watch/warning format that served as a control and four experimental formats that included some expression of tornado likelihood (described in

<sup>4</sup> These probabilities were based on tornado statistics from 12 weather forecast offices in the southeastern United States from 1 April 2014 through 31 October 2017 (C. Qin et al. 2021, unpublished manuscript).

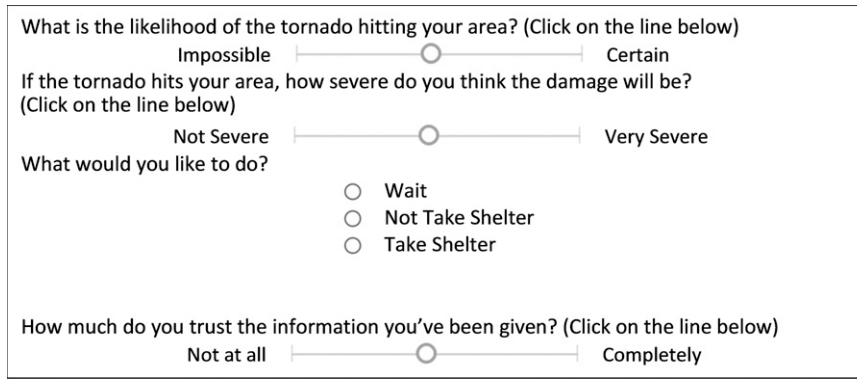


FIG. 3. Questions asked at each update on each trial and response mode: VAS.

detail below). The watch/warning format was similar to the communication currently in use. Participants were told that a “watch” meant that tornadoes were possible in or near the watch area, and a “warning” meant that a tornado was imminent or occurring and that taking shelter was advised. In this simplified task, a watch was issued if the storm reached a position at which the actual probability of hitting home was greater than or equal to 0.13 and less than or equal to 0.24. In this probability range, based on the cost–loss structure described above in Table 1, the economically optimal decision was to wait. Here, we define the optimal decision as the one for which the participant could expect to lose the least points either in terms of the cost or expected loss (no gains are possible in this scenario). This constitutes the operationalization of the quality or “goodness” or participants decisions. The expected loss was the potential penalty (1500 points) weighted by the probability of receiving it, the probability the tornado hitting home from that position. See Eqs. (A1)–(A3) in the appendix. For positions at which the actual probability of hitting home was less than 0.13, no watch or warning was issued. In this probability range the optimal decision was to not take shelter. A warning was issued for positions at which the actual probability of hitting home was 0.25 or greater, when the optimal decision was to take shelter.

The four experimental formats included some form of explicit uncertainty information conveying the underlying percent chance of the tornado hitting home from the current location of the storm (Fig. 5). In the color-coded condition participants received a color-coded forecast and were told in the instructions that warmer colors indicated higher likelihood of a tornado hitting their area. A green bar was shown when the actual probability of a tornado hitting home was less than 0.13, yellow when the actual probability was  $\geq 0.15 \leq 0.24$ , and orange when the actual probability was  $\geq 0.25$  and  $\leq 0.40$ . Notice that these values are identical to those defining “no watch or warning,” “watch,” and “warning” and correspond to the optimal decision thresholds.

In the percent chance condition participants saw a numeric percent chance of a tornado ranging from 0% to 40% on a continuous scale and rounded to two decimal places. In the percent chance range condition participants saw ranges 0%–12%, 13%–24%, and 25%–40% (shown in Fig. 5, column D). Participants in the color + percent chance range condition saw both the color bar and the corresponding percent chance range (e.g., green and 0%–12%). Each forecast in the experimental conditions was preceded with the text: “Chance of tornado in your area:” (e.g., 6%). Figure 6 shows the storm grid and describes the stimuli that were shown in each condition at each position. It is important to note that in conditions with categorized expressions (watch and warning, color, percent chance range), the category boundaries were aligned to the optimal likelihoods for each of the three decision choices.

d. Design

This experiment was a two-factor between/within-mixed design. The two independent variables were 1) forecast format (between groups) and 2) actual probability (within groups). The background information, task goal, cost–loss structure, and underlying probabilities were the same for all participants. The critical difference was the format by which the forecast information was presented (factor 1). Participants were randomly assigned<sup>5</sup> to one of five format conditions:

		Longitude							
		1	2	3	4	5	6	7	8
Latitude	1	0	0	0	0.0513	0.027	0	0	0
	2	0	0	0.1248	0.1188	0.108	0.09	0	0
	3	0	0.1812	0.1923	0.2064	0.225	0.24	0.3	0
	4	0.1905	0.2044	0.2226	0.247	0.28	0.34	0.4	1
	5	0	0.1812	0.1923	0.2064	0.225	0.24	0.3	0
	6	0	0	0.1248	0.1188	0.108	0.09	0	0
	7	0	0	0	0.0513	0.027	0	0	0

FIG. 4. Actual probability of tornado hitting home from each cell on the grid.

<sup>5</sup> Random assignment allowed us to assume that any differences in abilities or experience would be distributed across conditions.




A: Watch and Warning	B: Color	C: Percent Chance	D: Percent Chance Range	E: Color + Percent Chance Range
No watch or warning		6%	0 – 12%	B + C
Watch		19%	13 – 24%	B + C
Warning		33%	25 – 40%	B + C

FIG. 5. Forecasts by format (in the columns labeled A–E) and optimal decision threshold ranges. All forecasts were preceded by the phrase “Chance of Tornado Hitting Your Area.” The percentages shown in column C are the midpoints of the ranges of percentages (identical to that in column D) shown in that condition.

1) watch and warning, 2) color, 3) percent chance, 4) percent chance range, and 5) color + percent chance range. Therefore, each participant saw only a single format throughout the experiment (between groups). The second factor was the probability level: 1) 0%–12%, 2) 13%–24%, and 3) 25%–40%—all levels shown to all participants (within groups). There were four dependent variables: forecast understanding (operationalized using likelihood and severity ratings), decision quality (operationalized as expected loss), decision timeliness, and trust (rating) in the forecast.

### 3. Results

Our overarching question for this research was whether the differences in forecast format affected participants’ understanding of

the likelihood of a tornado, trust in the forecast, or their decisions. Therefore, the results are presented here in three parts, those addressing 1) understanding, 2) trust, and 3) decision-making (quality and timeliness). In each section, analyses of variance (ANOVA) were conducted assessing the effect of forecast format (watch and warning, color, color + percent chance range, percent chance range, and percent chance) on each dependent variable. ANOVA was selected because it is appropriate for categorical independent and continuous dependent variables. It allowed us to determine whether there were any systematic differences due to forecast format on the dependent variables tested here.

Effect sizes for ANOVAs were measured with partial eta-squared values. Effect sizes for contrasts were measured with Cohen’s *d*. Planned contrasts were corrected for familywise

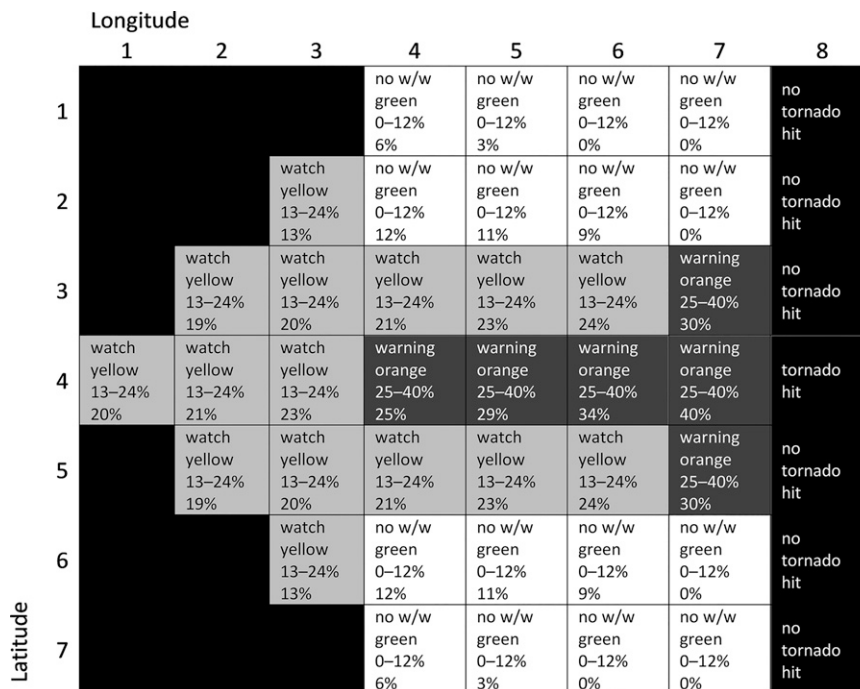


FIG. 6. Storm location grid, with stimuli shown in each forecast format. Optimal decisions are indicated by shading (white = not take shelter, light gray = wait, and dark gray = take shelter). Cells shaded black are impossible positions or are tornado destinations (column marked as 8).

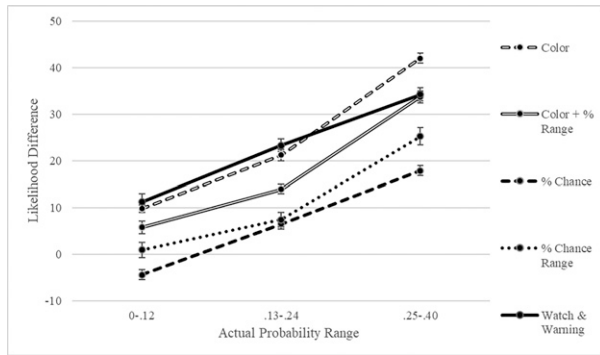


FIG. 7. Average likelihood difference, by actual probability range and forecast format.

error using Bonferroni correction ( $\alpha = 0.0125$ ). Post hoc comparisons were corrected for multiple comparisons using the Tukey test. [The hypotheses tested below were registered with the Open Science Framework (<https://osf.io/4uc67/>).] Open science framework is a tool that promotes openness and transparency in the research life cycle by hosting time-stamped registrations of research hypotheses, methods, and analysis plans (Foster and Deardorff 2017). Researchers post this information prior to conducting analyses.

a. Understanding of the forecast (hypothesis 1)

There were two issues related to understanding. The first was how closely participants' interpretation of likelihood, as indicated by their response on the VAS, matched the actual likelihood value and whether that varied by forecast format. The second was whether using some formats (e.g., color coding), participants mistook likelihood for severity.

1) LIKELIHOOD LEVEL

To determine how close participants' likelihood estimate was to the intended value, the actual probability of a tornado hitting home on a given update was subtracted from the participant's likelihood rating for that update. Participants likelihood rating was summarized as the percentage of the line between the left anchor (*impossible*) and the position to which participants moved the handle on the VAS. Then, in order to determine whether the error differed by actual probability level (e.g., higher for 25%–40% range), a mean

difference score was calculated for each participant in the three actual probability ranges, (i.e., 0%–12%, 13%–24%, and 25%–40%). Negative numbers indicate that participants' likelihood ratings were lower on average than the actual probability of a tornado hitting home, while positive numbers indicate that it was higher (see Fig. 7 and Table 2). Then, in order to determine whether estimates differed by condition, a mean was taken across participants in each forecast format condition, in each probability range (e.g., mean likelihood difference across participants in color-coded condition in 0%–12% probability range, color coded green).

Participants overestimated likelihood in all forecast formats but did so to the greatest degree with color-coded forecasts and at the highest actual probability range (Fig. 7; Table 2). A mixed model ANOVA was conducted on mean likelihood difference scores with actual probability range (0–0.12; 0.13–0.24; 0.25–0.40) as the within-groups factor and forecast format (watch and warning, color, color + percent chance range, percent chance range, and percent chance) as the between-groups factor. There was a significant main effect of forecast format  $F(4, 484) = 47.01, p < 0.001$ , with  $\eta_p^2 = 0.28$ . Planned contrasts, comparing the control (watch and warning) with experimental conditions revealed that the mean likelihood difference in the watch and warning condition was significantly greater than in all of the other conditions [color + percent chance range:  $t(484) = 3.30, p < 0.001$ , and Cohen's  $d = 0.30$ ; percent chance range:  $t(484) = 7.60, p < 0.001$ , and Cohen's  $d = 0.69$ ; and percent chance:  $t(484) = 10.30, p < 0.001$ , and Cohen's  $d = 0.94$ ] except the color only condition. Furthermore, Tukey post hoc comparisons showed that likelihood differences in both color:  $t(484) = 8.60, p < 0.001$ , and Cohen's  $d = 0.78$ , and color + percent chance range:  $t(484) = 4.30, p < 0.001$ , and Cohen's  $d = 0.39$ , conditions were significantly greater than in the percent chance range condition, suggesting that color significantly increased overestimation despite the addition of numbers. However, there was no significant difference between the percent chance range and percent chance conditions,  $p > 0.05$ , suggesting that ranges confer a similar advantage as individual probabilities.

There was also a significant main effect of actual probability range suggesting that overestimation increased in higher actual probability levels,  $F(1.53, 739.05) = 932.37, p < 0.001$ , with  $\eta_p^2 = 0.66$ . The degrees of freedom were corrected for violation of sphericity using the Greenhouse–Geisser correction. There was also a significant interaction between actual

TABLE 2. Mean differences (with standard deviations) between participants' likelihood rating and actual probability, by format forecast and actual probability range (0–0.12, 0.13–0.24, and 0.25–0.40). The final column shows the overall mean difference by forecast format.

Forecast format	0–0.12		0.13–0.24		0.25–0.40		Overall mean	
	M	SD	M	SD	M	SD	M	SD
Watch and warning	11.2	12.1	23.3	15.2	34.3	17.4	21.93	12.7
Color	9.76	12.3	21.2	13.8	42.0	19.2	21.89	11.95
Color + percent chance range	5.76	10.4	13.9	11.4	33.7	16.2	15.23	10.29
Percent chance range	0.91	10.3	70.43	10.7	25.3	14.6	80.89	90.8
Percent chance	–4.41	7.7	60.53	10.3	17.9	12.9	50.69	90.1



probability range and forecast format  $F(6.11, 739.05) = 8.07$ ,  $p < 0.001$ , with  $\eta_p^2 = 0.06$ . As shown in Fig. 7, likelihood difference for watch and warning and color were similar in the bottom two percent chance ranges. However, the overestimation for color was much greater in the upper range [mean  $M = 42.02$ ; standard deviation (SD) = 19.23] than for watch and warning ( $M = 34.26$ ; SD = 17.45).

In sum, forecasts that included a numeric estimate of likelihood were less susceptible to overestimation than were those that did not (color and watch and warning). Surprisingly, percent chance range performed in a manner that was comparable to percent chance alone, suggesting that understanding was substantially better not only with single probabilities but also with second-order numeric uncertainty information.

## 2) LIKELIHOOD–SEVERITY CONFUSION

We were also concerned that participants would misinterpret the expression of likelihood as an expression of severity, especially when color coding was used. Mistaking likelihood for severity was operationalized in three ways. The first was the difference between severity ratings and the likelihood ratings. The second was the correlation between severity and likelihood ratings. The third was the variability of severity ratings (in fact severity was held constant so less variability represents better understanding). There was evidence for this misunderstanding in all three operationalizations. Severity ratings were summarized as the percentage of the rating line between the left anchor (not severe) and the position to which participants moved the handle.

For the first operationalization, the likelihood rating was subtracted from the severity rating made at the same forecast update (severity – likelihood). Then an average difference score was calculated for each participant. The smaller the difference (positive or negative) was, the more similarly the participants regarded the two constructs and the greater was the confusion (see Fig. 8). An ANOVA conducted on mean difference score showed a main effect for forecast format  $F(4, 484) = 8.05$ ,  $p < 0.001$ , with  $\eta_p^2 = 0.06$ , suggesting that the difference was least (confusion greatest) in the color ( $M = 2.43$ ; SD = 10.45) and most (confusion least) in the percent chance range ( $M = 12.92$ ; SD = 22.46) condition. Planned contrasts revealed that the mean difference for color was significantly smaller (more confusion) than percent chance range:  $t(484) = 4.15$ ,  $p < 0.001$ , and Cohen's  $d = 0.38$ , and percent chance conditions ( $M = 13.14$ ; SD = 20.84):  $t(484) = 4.15$ ,  $p < 0.001$ , and Cohen's  $d = 0.38$ . Although the difference between color + percent chance range ( $M = 7.76$ ; SD = 17.75) and color approached significance,  $t(484) = 2.10$ ,  $p = 0.037$ , and Cohen's  $d = 0.19$ , it did not fall below the Bonferroni corrected level of 0.0125. Nor did the difference between color and watch and warning ( $M = 3.06$ ; SD = 15.88) reach significance. This suggests that forecasts that included a numeric estimate of likelihood were less susceptible to the likelihood–severity confusion than were those that did not. However, including color reduced this corrective effect to some degree and gave rise to confusion despite the presence of the numeric expression in the color + percent range condition.

Mistaking likelihood for severity was also operationalized as the correlation between severity and likelihood ratings.

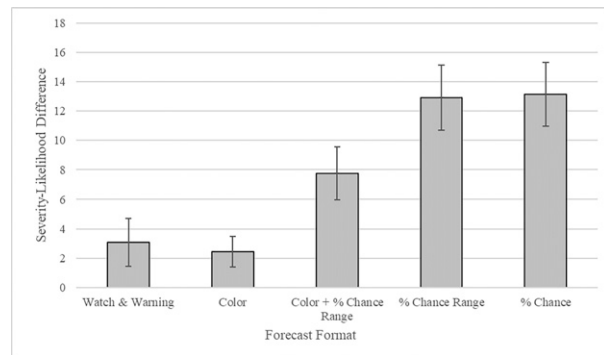


FIG. 8. Mean difference between severity and likelihood ratings, by forecast format.

Because the sampling distribution for highly correlated variables is skewed, the distributions were submitted to Fisher's  $z$  transformation before conducting the ANOVA. The ANOVA revealed a significant main effect of forecast format  $F(4, 484) = 6.36$ ,  $p < 0.001$ , with  $\eta_p^2 = 0.05$ , showing that correlations were highest in the color condition ( $M = 0.79$ ; SD = 0.30) and least in the percent chance range condition ( $M = 0.56$ ; SD = 0.43). Planned contrasts showed that the correlation in the color condition was significantly higher than in the percent chance range:  $t(484) = 4.50$ ,  $p < 0.001$ , and Cohen's  $d = 0.41$ , and percent chance conditions ( $M = 0.58$ ; SD = 0.46):  $t(484) = 4.08$ ,  $p < 0.001$ , and Cohen's  $d = 0.37$ . However, the correlation in the color condition did not significantly differ from the watch and warning ( $M = 0.69$ ; SD = 0.36):  $t(484) = 2.45$  and  $p = 0.015$ , and color + probability range ( $M = 0.68$ ; SD = 0.37):  $t(484) = 2.24$  and  $p = 0.025$ , conditions. This again suggests that forecasts that included color were most susceptible to the likelihood–severity confusion.

The third operationalization of the likelihood–severity confusion was the variability in severity ratings represented by the mean standard deviation of each participant's severity ratings. Recall that participants were informed that all storms involved same wind speed implying the same severity. Thus, their severity ratings should be the same throughout the experiment. Smaller mean standard deviation values indicated less variability and therefore less likelihood–severity confusion. A one-way ANOVA on mean severity SD revealed a significant main effect of forecast format  $F(4, 484) = 9.72$ ,  $p < 0.001$ , with  $\eta_p^2 = 0.07$  such that the SD was greatest in the color condition ( $M = 18.46$ ; SD = 7.07) and least in the percent chance ( $M = 13.46$ ; SD = 6.52) and percent chance range ( $M = 13.46$ ; SD = 6.08) conditions. Planned contrasts revealed that the severity SD of the color condition was significantly greater than all other conditions: watch and warning ( $M = 15.44$ ; SD = 6.56):  $t(484) = 3.20$ ,  $p = 0.001$ , and Cohen's  $d = 0.29$ ; color + percent chance range ( $M = 15.96$ ; SD = 6.87):  $t(484) = 2.66$ ,  $p = 0.008$ , and Cohen's  $d = 0.24$ ; percent chance range ( $M = 13.46$ ; SD = 6.08):  $t(484) = 5.38$ ,  $p < 0.001$ , and Cohen's  $d = 0.49$ ; and percent chance conditions ( $M = 13.46$ ; SD = 6.52):  $t(484) = 5.25$ ,  $p < 0.001$ , and Cohen's  $d = 0.47$ . This suggests that participants in the

color-coded condition were confusing severity and likelihood to a greater degree than those using other formats.

Thus, analyses of all three operationalizations of this misunderstanding (severity difference, severity–likelihood correlation, severity standard deviation) suggest that color coding tends to promote, to a greater degree than the other formats tested here, a confusion between likelihood and severity. Participants using color-coded likelihood expressions tended to think that they were also receiving information about severity. Including a numeric expression (color + percent chance range) appears to counteract this tendency to some degree as shown in two of the three operationalizations, but not completely.

### b. Trust in the forecast (hypothesis 2)

To determine whether explicit numeric likelihood estimates inspired greater trust, we examined trust in the forecast rated after learning the outcome of the storm at the end of each trial. Trust ratings were summarized as the percentage of the rating line between the left anchor, “not at all,” and the position to which participants moved the handle. An ANOVA conducted on mean postoutcome trust rating revealed a main effect of forecast format  $F(4, 484) = 3.10, p = 0.016$ , with  $\eta_p^2 = 0.025$  such that trust was highest in the percent chance condition ( $M = 50.38$ ;  $SD = 18.80$ ) and lowest in the color condition ( $M = 42.82$ ;  $SD = 15.75$ ). Tukey post hoc comparisons showed that trust ratings for percent chance was significantly greater than for color:  $t(484) = 2.91, p = 0.031$ , and Cohen’s  $d = 0.26$ . In addition, trust was significantly higher in the color + percent chance range [ $M = 50.07$ ;  $SD = 19.00, t(484) = 2.82, p = 0.040$ , and Cohen’s  $d = 0.26$ ] than in the color condition. No other significant differences were found.

In sum, the percent chance forecast inspired the greatest trust. Color, which inspired the least trust, also gave rise to the greatest overestimation of likelihood, which may explain the low trust. Because the forecasts were well calibrated, overestimation of likelihood may have led participants to expect more tornado hits than they actually experienced. However, adding the percent chance range to color, helped to preserve trust. We return to this issue in the discussion section below.

### c. Decision-making (hypothesis 3)

There were three issues related to decision-making that were of interest, decision quality, cautiousness, and timeliness. We hypothesized that both quality and timeliness would improve with numeric likelihood estimates. We had no specific hypotheses about cautiousness.

#### 1) DECISION QUALITY

To determine the effect of forecast format on decision quality, it was operationalized as the expected loss/cost of participants’ final decisions (wait, take shelter, or not take shelter). Each decision was assigned either a point cost (take shelter) or the expected point loss of a penalty (wait or not take shelter). The expected loss can be thought of as the penalty amount weighted by the chance that it would be incurred

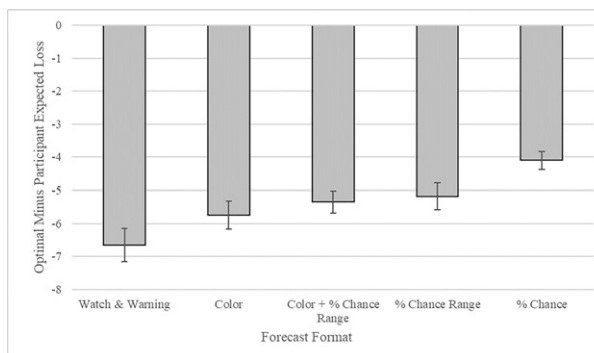


FIG. 9. Mean expected loss difference, by forecast format. Means closer to zero indicate better decisions, because there is little difference between the participant’s and the optimal expected loss values.

(probability of the tornado hitting home). See appendix for the expected loss calculations. The cost or expected loss of the participants’ decision was then subtracted from the optimal expected value (or cost) on that trial. The smaller the expected loss difference the better the participants decision.

The ANOVA conducted on expected loss difference revealed a main effect for forecast format  $F(4, 484) = 5.19, p < 0.001$ , with  $\eta_p^2 = 0.041$  such that participants in the percent chance condition had the smallest expected loss difference ( $M = -4.10$ ;  $SD = 2.70$ ), and those in the watch and warning condition had the largest ( $M = -6.65$ ;  $SD = 4.92$ ) (see Fig. 9). Planned contrasts revealed that decision quality in the percent chance:  $t(484) = -4.40, p < 0.001$ , and Cohen’s  $d = 0.40$  and the percent chance range ( $M = -5.18$ ;  $SD = 4.11$ ):  $t(484) = 2.62, p = 0.009$ , and Cohen’s  $d = 0.21$ , conditions were significantly better (smaller difference) than in the watch and warning condition. Tukey post hoc comparisons revealed that percent chance ( $M = -4.10$ ;  $SD = 2.70$ ) was also significantly less (better) than color alone ( $M = -5.75$ ;  $SD = 4.31$ ):  $t(484) = 2.89, p = 0.032$ , and Cohen’s  $d = 0.26$ . No other significant differences were found. This suggests that the forecast showing the percent chance of a tornado led to better decisions than the conventional watch and warning as well as color coded likelihood.

#### 2) DECISION CAUTIOUSNESS

Although we had no specific hypotheses about this issue, it was important to evaluate cautiousness. It could be that some forecast formats (e.g., watch/warning) lead to more cautious decisions without leading to economically optimal decisions. That is because greater cautiousness means shelter more often overall, whereas economically optimal means sheltering only when that option constitutes the least point loss. To understand whether participants were more cautious (i.e., shelter more often) with certain forecast formats, a one-factor ANOVA was conducted on the proportion of take-shelter decisions (across the 40 trials regardless of whether take shelter was the optimal choice) with the independent variable forecast format (watch and warning; color; color + percent chance

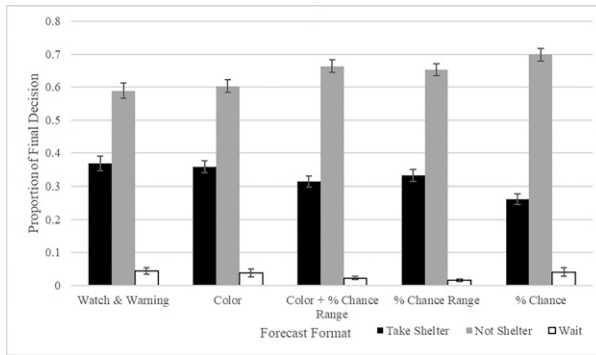


FIG. 10. Proportion of final decision out of 40, by forecast format.

range; percent chance range; percent chance; see Fig. 10). There was a significant main effect of forecast format  $F(4, 484) = 5.19, p < 0.001$ , with  $\eta_p^2 = 0.041$ . Participants with the watch and warning [ $M = 0.37, SD = 0.21, t(484) = 4.04, p < 0.001$ , and Cohen's  $d = 0.37$ ] and color conditions [ $M = 0.36, SD = 0.18, t(484) = 3.73, p = 0.002$ , and Cohen's  $d = 0.34$ ] had a significantly higher proportion of shelter decisions than those with the percent chance condition ( $M = 0.26; SD = 0.16$ ). Thus, although participants in the watch and warning and color conditions made fewer good decisions from an economic standpoint, they were more cautious overall.

### 3) DECISION TIMELINESS

The final analysis was conducted on decision timeliness, operationalized as the difference in the forecast update number (1–7) at which the participant made a final decision (shelter or not take shelter) and the point at which it was optimal to make a final decision (stopping difference). It was optimal to make a final decision on the first trial on which the cost of sheltering or the expected loss of not taking shelter was the less than waiting. The step at which the participant made a final decision was subtracted from the optimal stopping point on each trial. A negative number means that the participant made a decision prior to the optimal stopping point and a positive number means the final decision was made after the optimal stopping point; zero indicates that the two are the same (optimal). A mean was taken for each participant across all 40 trials.

Participants on average made timely decisions in all forecast formats although those in the percent chance condition showed a slight delay beyond optimal stopping (see Fig. 11). A one-factor ANOVA conducted on stopping difference revealed a significant main effect for forecast format  $F(4, 484) = 2.45, p = 0.046$ , with  $\eta_p^2 = 0.02$ . The stopping difference in the percent chance condition ( $M = 0.29; SD = 1.22$ ) was significantly different than watch and warning ( $M = -0.33; SD = 1.81; t(484) = 2.71, p = 0.007$ , and Cohen's  $d = 0.25$ ). No other significant differences were found. Thus, although percent chance showed a slight delay, those in the watch and warning condition made their decisions slightly too early by about the same amount (about  $1/3$  of a step).

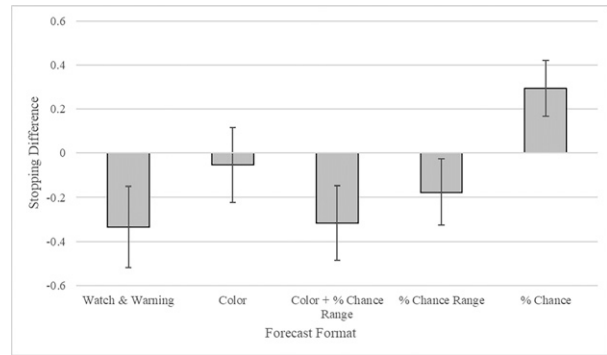


FIG. 11. Mean difference in participants' final decision and optimal stopping point, by forecast format.

In sum, participants in the percent chance condition waited slightly longer to make a decision; however, they made decisions with the highest expected value. Percent chance range showed comparable decision quality to percent chance alone but did not delay significantly beyond optimal stopping. Participants in the watch and warning and color conditions made more cautious decisions and earlier, although their decisions had the greatest expected loss.

## 4. Discussion

This experiment provides strong evidence for the benefits of numeric likelihood information to communicate forecast uncertainty in a dynamic decision environment in which multiple sequential forecasts must be evaluated to make a final decision. These benefits were seen in terms of user's understanding of and trust in the forecast as well as in the decisions that they made based on the forecasts.

### a. Understanding

Participants in the conditions in which numerical expressions alone were provided (percent chance and percent chance range) understood the forecast best both in terms of the intended likelihood and avoiding the potential confusion with severity. However, the benefits of color-coded likelihood are less clear.

#### 1) OVERESTIMATING LIKELIHOOD

Although participants using all forecast formats showed some tendency to overestimate the likelihood of a tornado hitting their area, those using forecasts that included color coding overestimated to the greatest degree. This is in line with previous research showing that warnings with color are perceived as more hazardous overall (Braun et al. 1995). Moreover, in the study reported here, the bias increased with the actual probability. Although the overestimation was only about 10% at the low end, it increased to 42% overestimation at the high end, more than doubling the intended amount. This was despite the fact that the upper range shown in the color coded condition was not red (often ranked as indicating the greatest risk) but rather orange. A possible explanation in

the color only condition, which did not include a numeric key, is that participants assumed that each color category represented one-third of a scale from *impossible* (0%) to *certain* (100%). Under this assumption, the mean likelihood estimates in the 0–0.12, 0.13–0.24, and 0.25–0.40 actual probability ranges would have been mistaken as 0%–33%, 34%–66%, and 67%–100%, respectively. Although participants' mean likelihood estimates in the color condition fall within the bottom two of the inflated ranges, 0%–33% ( $M = 29.0$ ) and 34%–66% ( $M = 40.5$ ) they were much lower ( $M = 61.25$ ) than would be expected in the top range, 67%–100%, ruling out this explanation. Moreover, it is important to note that the increase in overestimation extended to the color-coded condition that *did* include a numeric key (color + percent chance range), although the overestimation was not as great as with color alone. An alternative explanation is that color coding led to greater arousal, which has been shown to inflate perceived likelihood (Vosgerau 2010) and may have increased with warmer colors. This is something that could be explored in future research by adding a measure of arousal. In sum, although the exact mechanism for the effect is unclear, it is clear that color coding introduces a bias in perceived likelihood that could be problematic in some situations. Users, inflating the likelihood, may come to regard color-coded expressions as untrustworthy over time when the frequency of the outcome is less than expected. Adding a key explaining the numeric range represented by each color attenuated but did not completely eliminate the bias.

Participants also significantly overestimated the likelihood in the watch and warning condition, although probably for different reasons. The watch and warning format that functioned as the control did not provide overt likelihood information. However, by definition, a tornado warning means that a tornado is imminent or occurring and thus, may have implied 100% certainty rather than the 25%–40% that was the case here. Again, however, the mean estimate in the warning condition did not approach 100% ( $M = 34.3$ ;  $SD = 17.4$ ). Thus, although participants anticipated some uncertainty even in the watch and warning condition, they may have assumed that the threshold for a warning was higher than it actually was, which would account for the overestimation in likelihood observed here.

## 2) LIKELIHOOD–SEVERITY CONFUSION

In addition, participants had a tendency to misinterpret color-coded likelihood information as indicating something about severity. This was supported in analyses of all three operationalizations of the likelihood–severity confusion (severity difference, severity–likelihood correlation, and severity SD) and observed despite the fact that participants were explicitly told that the colors indicated the chance of a tornado. This misinterpretation may be due to previous exposure to color-coded weather charts intended to communicate severity, such as amount of precipitation or intensity of wind speed. Perhaps prior exposure established expectations about the meaning of color coding, that color generally indicates something about severity, which were not fully contradicted by

explanations provided in context of the task. Under some circumstances the likelihood–severity confusion may be beneficial, serving to encourage caution. However, it may also damage trust when expectations are not confirmed. This misinterpretation could be particularly dangerous when the magnitude of the storm is great, but the likelihood is low. Residents may interpret the color to mean that if a tornado were to occur, the damage would be minimal. This may, in turn, reduce their willingness to take protective action. It is important to note that misinterpreting likelihood as severity was least in the numeric conditions in which color was not used.

The likelihood–severity confusion was also observed in the watch and warning condition, for which the explanation may be slightly different. In the absence of explicit likelihood information (as in the watch and warning condition) one might be unable to disentangle the concepts of likelihood and severity and instead assume that as one increases so does the other. Indeed, there is some evidence that verbal descriptions of likelihood are translated into a higher percent chance when the outcomes are severe as opposed to neutral, referred to as severity bias (Harris et al. 2009; Weber and Hilton 1990). Importantly, in the experiment reported here, participants in the percent chance and percent chance range conditions were least susceptible to the likelihood–severity confusion. This aligns with research suggesting that numeric expressions of probability tend to reduce the severity bias (Fischer and Jungermann 1996).

### b. Trust

The percent chance forecast also inspired the greatest trust. Color, which gave rise to the greatest overestimation of likelihood, inspired the least trust. However, adding the percent chance range to color helped to preserve trust despite inflated likelihood estimates, mirroring the positive impact of uncertainty estimates on trust seen in previous research (LeClerc and Joslyn 2015).

### c. Decision-making

Not only did numeric expressions of uncertainty lead to better understanding, but they also led to better decisions in this dynamic environment. Participants with numeric likelihood forecasts made decisions with higher expected value than those using categorical forecasts (i.e., color coded or watch and warning forecast). This was true even though each forecast included multiple updates, each with its own probabilistic estimate and each choice point included three options (wait, take shelter, or not take shelter). In sum, these results suggest that participants were able to evaluate and incorporate numeric uncertainty information despite the increase in processing load represented by these complexities.

The advantage of percent chance over categorical expressions is particularly striking in this experimental context because the boundaries of the categories were aligned with the optimal decision. When the economically optimal decision was to not take shelter, the color was green and no warning was shown. When the economically optimal decision was to

wait, the color was yellow and a watch was shown. When the economically optimal decision was to take shelter, the color was orange and a warning was shown. Therefore, those in the categorical conditions had information that was perfectly tailored to the cost–loss structure of the task. If participants had chosen to shelter only when the highest category was forecast (i.e., warning, or orange), the expected value in these conditions would have been superior (i.e., no difference from optimal). This was not the case. Instead, participants in both the color-coded and watch and warning conditions chose to take precautionary action more frequently, often unnecessarily expending resources to avoid a tornado that was unlikely.

While the percent chance expression proved to be advantageous overall, there was a slight delay in making the decision (timeliness), less than a single forecast update on average, which may be the only indication of the extra processing required with numeric expressions of uncertainty. In addition, those using the percent chance expression, although they made better decisions from an economic standpoint, were not as cautious as those using the watch and warning and color-coded forecasts who took precautionary action more often overall. Thus, there might be a trade-off in the choice of forecast format between precision and cautiousness, such that some forecast formats enhance the first while others the second. Some have argued that such trade-offs could be exploited to fit specific situations (Ash et al. 2014). For instance, color coding might be used in situations in which additional cautiousness would be beneficial.

It is important to note, however, that the increased cautiousness observed here is likely directly due to misunderstanding the forecast (as reflected in likelihood ratings). As such, there may be a cost in terms of lowered trust that could affect responses to subsequent warnings. Indeed, postoutcome trust was significantly lower in the color-coded than in the percent chance condition. It may be that trust in color-coded forecasts was diminished because of the frequent and ultimately unnecessary decisions to take shelter, suggesting a “false alarm” effect, in which participants saw themselves as receiving an inaccurate forecast.

#### *d. Second-order uncertainty*

Importantly, the advantages for numeric uncertainty expressions extended to the percent chance range condition, which was also categorical in a sense and added a level of uncertainty about the likelihood itself. Nonetheless, those using the percent chance range had a better understanding of the likelihood of a tornado than did those using the watch and warning as well as those using color + percent chance range, which was identical with the exception of color. Those using the percent chance range were also less susceptible to the likelihood–severity confusion than those using color alone (as seen in the severity standard deviation and severity difference operationalizations). Trust in this format was equal to that of the single percent chance expression, despite the fact that it that included uncertainty about the probabilistic forecast itself. Moreover, decision quality in percent chance range was equivalent to percent chance while decisions were made in a

timelier fashion. Taken together, these results suggest that in a real-world settings, forecasters may be able to include second-order uncertainty information (e.g., a range such as 0%–12%), without a detriment to understanding, trust or decision quality.

#### *e. Limitations*

As an experimental study using a convenience sample of college students there are several limitations that should be mentioned. First, as with all experimental studies there was a trade-off between ecological validity and experimental control. In a natural environment several other factors would also influence peoples’ decisions. Here those factors have been stripped away or held constant in order to determine the impact of forecast format alone. Indeed, research using more realistic storm-tracking tasks suggests that with multiple potential information sources, search strategies must be adjusted (Wu et al. 2015a) and decisions can be delayed (Wu et al. 2015b).

In addition, the consequences in the real-world counterpart of this task would be far more serious. Here the consequences were merely a reduction in points that impacted the monetary reward. Although it is important to note that, unlike many laboratory-based studies, there were actual consequences in the task reported here. Participants who made better decisions received a larger cash bonus.

The college student participants may be more prepared to understand probabilistic forecasts than some members of the public, giving rise to questions about generalizability. However, recent evidence suggests that, although they may not have a theoretical understanding of probability, those with high school education or less can benefit from forecasts that include numeric likelihood estimates to the same degree as those who are college educated (Grounds and Joslyn 2018; Grounds et al. 2017).

#### *f. Next steps*

Although these are promising results, future research should replicate them using a more diverse sample to verify that generalization warranted. Research should also explore decision tasks with different kinds of complexities, such as multiple sources of information or multiple decision alternatives, to determine whether the advantages for numeric uncertainty estimates continue to hold. Finally, it would be useful to know if decision quality increases further with a combination of both numeric uncertainty information and explicit warning information in this complex scenario, as has been observed in simplified scenarios in previous research (Joslyn and LeClerc 2012).

## **5. Conclusions**

Taken together, the results of the study reported here provide compelling evidence for the benefits of including numeric uncertainty information in warning forecasts in a dynamic decision environment. Granted there are several other factors that influence protective action decisions in a real-world situation (Lindell 2018) that interact in complex ways to influence peoples’ decisions. Here we have examined one critical

TABLE A1. Storm location grid with expected loss of decisions. Here, W = wait, S = shelter, and NS = not take shelter. Columns represent longitude (1–8). Rows represent latitude (1–7).

	1	2	3	4	5	6	7	8
1	W: 0 S: 303 NS: 0	W: 56 S: 312 NS: 0	W: 107 S: 327 NS: 0	W: 97 S: 348 NS: 77	W: 81 S: 395 NS: 61	W: 60 S: 448 NS: 40	W: 80 S: 507 NS: 60	Tornado destination
2	W: 80 S: 303 NS: 0	W: 159 S: 312 NS: 0	W: 187 S: 327 NS: 187	W: 192 S: 348 NS: 178	W: 194 S: 395 NS: 182	W: 194 S: 448 NS: 175	W: 80 S: 507 NS: 60	Tornado destination
3	W: 196 S: 303 NS: 0	W: 265 S: 312 NS: 272	W: 281 S: 327 NS: 288	W: 308 S: 348 NS: 310	W: 336 S: 395 NS: 358	W: 373 S: 448 NS: 400	W: 530 S: 507 NS: 510	Tornado destination
4	W: 278 S: 303 NS: 286	W: 298 S: 312 NS: 307	W: 324 S: 327 NS: 334	W: 360 S: 348 NS: 371	W: 403 S: 395 NS: 440	W: 507 S: 448 NS: 550	W: 680 S: 507 NS: 660	Tornado destination
5	W: 196 S: 303 NS: 0	W: 265 S: 312 NS: 272	W: 281 S: 327 NS: 288	W: 308 S: 348 NS: 310	W: 336 S: 395 NS: 358	W: 373 S: 448 NS: 400	W: 530 S: 507 NS: 510	Tornado destination
6	W: 80 S: 303 NS: 0	W: 159 S: 312 NS: 0	W: 187 S: 327 NS: 187	W: 192 S: 348 NS: 178	W: 194 S: 395 NS: 182	W: 194 S: 448 NS: 175	W: 80 S: 507 NS: 60	Tornado destination
7	W: 0 S: 303 NS: 0	W: 56 S: 312 NS: 0	W: 107 S: 327 NS: 0	W: 97 S: 348 NS: 77	W: 81 S: 395 NS: 61	W: 60 S: 448 NS: 40	W: 80 S: 507 NS: 60	Tornado destination

component in that complex process in relative isolation, updating warning messages. Nonetheless, we believe that the contribution of this work to the bigger picture is key. It adds to the growing literature that nonexperts have at least a “working understanding” of fairly complex scientific information and can use it to their benefit. Although earlier evidence for the benefits of numeric uncertainty estimates (Joslyn and Leclerc 2013) focused on simple situations, here the advantage was seen in a more complex decision scenario in which participants received seven updates for the same event, each with its own probabilistic forecast, and three options at every decision point. Nonetheless, several clear advantages were seen for both percent chance of a tornado as well as percent chance range formats. These results have important implications for risk communication in the context of weather, water management and climate as well as other situations in which people need to make decisions about protection in uncertainty circumstance: People can understand fairly complex scientific information and make good use of the additional precision to improve their decisions, as long as the information is relevant and presented in an understandable format. Omitting explicit likelihood information when it is available not only leaves room for miscommunication between scientists/public officials and members of the public, but it may deprive decision-makers of information that could help them to make better choices.

*Acknowledgments.* We have no known conflicts of interest to disclose. This research was supported by a grant from the National Science Foundation DRMS:1559126.

*Data availability statement.* The data that support the findings of this study are available on request from the corresponding author.

## APPENDIX

### Expected Loss Calculations

If the participant chose to shelter, there was a one-time cost (see Table 1). If the participant chose to not shelter or to wait, the expected loss depended on the actual probability of the tornado hitting home at that particular storm position. The expected loss to not take shelter ( $EL_{\text{not\_shelter}}$ ) was the product of 1500-point hit penalty and the probability of experiencing that penalty (i.e., a tornado hit home) plus any wait cost that had been incurred prior to the final decision:

$$EL_{\text{not\_shelter}} = \text{Probability of tornado hitting home} \times 1500 + \text{cost}_{\text{wait}} \quad (\text{A1})$$

The expected loss of a wait decision ( $EL_{\text{wait}}$ ) was similar but also depended on the longitude at which the decision was being made. For all cells in longitude 7, it was the product of the penalty (1500) and probability of the tornado hitting home at that storm position plus the 80-point prior wait cost (20 points per wait decision for longitudes 3–6):

$$\text{Longitude 7: } EL_{\text{wait}} = 80 \text{ points incurred cost}_{\text{wait}} + (\text{probability of a tornado hitting home} \times 1500). \quad (\text{A2})$$

The expected loss of wait at longitude 6 (and all previous longitudes) depended on the three adjacent cells to the east (see section 2b). The smallest expected loss ( $EL_{\text{min}}$ ) in each of those cells was multiplied by the probability of the tornado moving to that cell from its current position. The sum of these three products was regarded as the expected value of waiting in the current location:

$$\text{Longitude 1: } EL_{\text{wait}}(\text{LAT}, \text{LON}) = 0.7 \times EL_{\text{min}}(\text{LAT}, \text{LON} + 1) + 0.3 \times EL_{\text{min}}(\text{LAT} + 1, \text{LON} + 1),$$

$$\begin{aligned} \text{Longitude 2-6: } EL_{\text{wait}}(\text{LAT}, \text{LON}) &= 0.3 \times EL_{\text{min}}(\text{LAT} - 1, \text{LON} + 1) + 0.4 \times EL_{\text{min}}(\text{LAT}, \text{LON} + 1) \\ &+ 0.3 \times EL_{\text{min}}(\text{LAT} + 1, \text{LON} + 1), \quad \text{and} \end{aligned}$$

$$\text{Longitude 7: } EL_{\text{wait}}(\text{LAT}, \text{LON}) = 0.7 \times EL_{\text{min}}(\text{LAT}, \text{LON} + 1) + 0.3 \times EL_{\text{min}}(\text{LAT} - 1, \text{LON} + 1). \quad (\text{A3})$$

At each storm position, the optimal decision was the one with the lowest cost or expected loss ( $EL_{\text{min}}$ ). A difference between the expected loss of the participants' decision and the optimal decision was calculated for each of their decisions  $i$  prior to the final decision  $n$ , and a mean was calculated for each trial ( $EL_{\text{difference}}$ ).<sup>A1</sup> Then, a mean was calculated for participants over the 40 trials and in each condition. Table A1 shows the expected loss values of all possible decisions by storm position. The participant's expected loss difference was a negative value or zero when the participant made the optimal decision:

$$EL_{\text{difference}} = \frac{\sum_{i=1}^n EL_{\text{min}} - EL_{\text{decision}}}{n}. \quad (\text{A4})$$

## REFERENCES

- Ash, K. D., R. L. Schumann III, and G. C. Bowser, 2014: Tornado warning trade-offs: Evaluating choices for visually communicating risk. *Wea. Climate Soc.*, **6**, 104–118, <https://doi.org/10.1175/WCAS-D-13-00021.1>.
- Baddeley, A., 2000: The episodic buffer: A new component of working memory? *Trends Cognit. Sci.*, **4**, 417–423, [https://doi.org/10.1016/S1364-6613\(00\)01538-2](https://doi.org/10.1016/S1364-6613(00)01538-2).
- Barnes, L. R., E. C. Grunfest, M. H. Hayden, D. M. Schultz, and C. Benight, 2007: False alarms and close calls: A conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140–1147, <https://doi.org/10.1175/WAF1031.1>.
- Borade, A. B., S. V. Bansod, and V. R. Gandhewar, 2008: Hazard perception based on safety words and colors: An Indian perspective. *Int. J. Occup. Saf. Ergon.*, **14**, 407–416, <https://doi.org/10.1080/10803548.2008.11076777>.
- Braun, C. C., P. B. Mine, and N. C. Silver, 1995: The influence of color on warning label perceptions. *Int. J. Ind. Ergon.*, **15**, 179–187, [https://doi.org/10.1016/0169-8141\(94\)00036-3](https://doi.org/10.1016/0169-8141(94)00036-3).
- Broad, K., A. Leiserowitz, J. Weinkle, and M. Stekete, 2007: Misinterpretations of the “cone of uncertainty” in Florida during the 2004 hurricane season. *Bull. Amer. Meteor. Soc.*, **88**, 651–668, <https://doi.org/10.1175/BAMS-88-5-651>.
- Chapanis, A., 1994: Hazards associated with three signal words and four colours on warning signs. *Ergonomics*, **37**, 265–275, <https://doi.org/10.1080/00140139408963644>.
- Cheong, L., S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham, 2016: Evaluating the impact of visualization of wildfire hazard upon decision making under uncertainty. *Int. J. Geogr. Inf. Sci.*, **30**, 1377–1404, <https://doi.org/10.1080/13658816.2015.1131829>.
- Dieckmann, N. F., R. Mauro, and P. Slovic, 2010: The effects of presenting imprecise probabilities in intelligence forecasts. *Risk Anal.*, **30**, 987–1001, <https://doi.org/10.1111/j.1539-6924.2010.01384.x>.
- Earle, T. C., 2010: Trust in risk management: A model-based review of empirical research. *Risk Anal.*, **30**, 541–574, <https://doi.org/10.1111/j.1539-6924.2010.01398.x>.
- Eiser, J. R., A. Bostrom, I. Burton, D. M. Johnston, J. McClure, D. Paton, J. van der Pligt, and M. P. White, 2012: Risk interpretation and action: A conceptual framework for responses to natural hazards. *Int. J. Disaster Risk Reduct.*, **1**, 5–16, <https://doi.org/10.1016/j.ijdrr.2012.05.002>.
- Fischer, K., and H. Jungermann, 1996: Rarely occurring headaches and rarely occurring blindness: Is rarely=rarely? The meaning of verbal frequentistic labels in specific medical contexts. *J. Behav. Decis. Making*, **9**, 153–172, [https://doi.org/10.1002/\(SICI\)1099-0771\(199609\)9:3<153::AID-BDM222>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1099-0771(199609)9:3<153::AID-BDM222>3.0.CO;2-W).
- Foster, E. D., and A. Deardorff, 2017: Open Science Framework (OSF). *J. Med. Libr. Assoc.*, **105**, 203–206, <https://doi.org/10.5195/jmla.2017.88>.
- Griffith, L. J., and S. D. Leonard, 1997: Association of colors with warning signal words. *Int. J. Ind. Ergon.*, **20**, 317–325, [https://doi.org/10.1016/S0169-8141\(96\)00062-5](https://doi.org/10.1016/S0169-8141(96)00062-5).
- Grounds, M. A., and S. L. Joslyn, 2018: Communicating weather forecast uncertainty: Do individual differences matter? *J. Exp. Psychol. Appl.*, **24**, 18–33, <https://doi.org/10.1037/xap0000165>.
- , S. Joslyn, and K. Otsuka, 2017: Probabilistic interval forecasts: An individual differences approach to understanding forecast communication. *Adv. Meteor.*, **2017**, 3932565, <https://doi.org/10.1155/2017/3932565>.
- Harris, A. J. L., A. Corner, and U. Hahn, 2009: Estimating the probability of negative events. *Cognition*, **110**, 51–64, <https://doi.org/10.1016/j.cognition.2008.10.006>.
- Hellier, E., M. Tucker, N. Kenny, A. Rowntree, and J. Edworthy, 2010: Merits of using color and shape differentiation to improve the speed and accuracy of drug strength identification on over-the-counter medicines by laypeople. *J. Patient Saf.*, **6**, 158–164, <https://doi.org/10.1097/PTS.0b013e3181eee157>.
- Joslyn, S., and S. Savelli, 2010: Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteor. Appl.*, **17**, 180–195, <https://doi.org/10.1002/met.190>.
- , and J. LeClerc, 2012: Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.*, **18**, 126–140, <https://doi.org/10.1037/a0025185>.

<sup>A1</sup> Because a participant's series of wait decisions and final decision may include cumulative wait costs, we take an average over decisions (rather than a sum) to prevent overestimation of wait costs over the trial.

- , and —, 2013: Decisions with uncertainty: The glass half full. *Curr. Dir. Psychol. Sci.*, **22**, 308–315, <https://doi.org/10.1177/0963721413481473>.
- , and S. Savelli, 2021: Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. *Front. Comput. Sci.*, **2**, 590232, <https://doi.org/10.3389/fcomp.2020.590232>.
- , K. Pak, D. Jones, J. Pyles, and E. Hunt, 2007: The effect of probabilistic information on threshold forecasts. *Wea. Forecasting*, **22**, 804–812, <https://doi.org/10.1175/WAF1020.1>.
- , L. Nadav-Greenberg, M. U. Taing, and R. M. Nichols, 2009: The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Appl. Cognit. Psychol.*, **23**, 55–72, <https://doi.org/10.1002/acp.1449>.
- Kahneman, D., and S. Frederick, 2002: Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 1st ed. T. Gilovich, D. Griffin, and D. Kahneman, Eds., Cambridge University Press, 49–81, <https://doi.org/10.1017/CBO9780511808098.004>.
- Karstens, C. D., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, <https://doi.org/10.1175/WAF-D-14-00163.1>.
- LeClerc, J., and S. Joslyn, 2015: The cry wolf effect and weather-related decision making. *Risk Anal.*, **35**, 385–395, <https://doi.org/10.1111/risa.12336>.
- Lesch, M. F., P.-L. P. Rau, Z. Zhao, and C. Liu, 2009: A cross-cultural comparison of perceived hazard in response to warning components and configurations: US vs. China. *Appl. Ergon.*, **40**, 953–961, <https://doi.org/10.1016/j.apergo.2009.02.004>.
- Lim, J. R., B. F. Liu, and M. Egnoto, 2019: Cry wolf effect? Evaluating the impact of false alarms on public responses to tornado alerts in the southeastern United States. *Wea. Climate Soc.*, **11**, 549–563, <https://doi.org/10.1175/WCAS-D-18-0080.1>.
- Lindell, M. K., 2018: Communicating imminent risk. *Handbook of Disaster Research*, 2nd ed. H. Rodríguez, W. Donner, and J. E. Trainor, Eds., Springer, 449–477.
- , and R. W. Perry, 2012: The protective action decision model: Theoretical modifications and additional evidence. *Risk Anal.*, **32**, 616–632, <https://doi.org/10.1111/j.1539-6924.2011.01647.x>.
- MacEachren, A. M., R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan, 2012: Visual semiotics & uncertainty visualization: An empirical study. *IEEE Trans. Vis. Comput. Graph.*, **18**, 2496–2505, <https://doi.org/10.1109/TVCG.2012.279>.
- Mayhorn, C. B., M. S. Wogalter, J. L. Bell, and E. F. Shaver, 2004: What does code red mean? *Ergon. Des.*, **12**, 12–14, <https://doi.org/10.1177/106480460401200404>.
- Miran, S. M., C. Ling, J. J. James, A. Gerard, and L. Rothfus, 2017: User perception and interpretation of tornado probabilistic hazard information: Comparison of four graphical designs. *Appl. Ergon.*, **65**, 277–285, <https://doi.org/10.1016/j.apergo.2017.06.016>.
- Nagele, D. E., and J. E. Trainor, 2012: Geographic specificity, tornadoes, and protective action. *Wea. Climate Soc.*, **4**, 145–155, <https://doi.org/10.1175/WCAS-D-11-00047.1>.
- NWS, 2011: False alarm reduction research. NOAA, [https://www.weather.gov/bmx/research\\_falsealarms](https://www.weather.gov/bmx/research_falsealarms).
- , 2012: Watch/warning/advisory definitions. NOAA, <https://www.weather.gov/lwx/WarningsDefined>.
- Radford, L., J. C. Senkbeil, and M. Rockman, 2013: Suggestions for alternative tropical cyclone warning graphics in the USA. *Disaster Prev. Manage.*, **22**, 192–209, <https://doi.org/10.1108/DPM-06-2012-0064>.
- Rashid, R., and M. S. Wogalter, 1997: Effects of warning border color, width, and design on perceived effectiveness. *Advances in Occupational Ergonomics and Safety II*, 1st ed. B. Das and W. Karwowski, Eds., IOS Press, 455–458.
- Ripberger, J. T., C. L. Silva, H. C. Jenkins-Smith, D. E. Carlson, M. James, and K. G. Herron, 2015: False alarms and missed events: The impact and origins of perceived inaccuracy in tornado warning systems. *Risk Anal.*, **35**, 44–56, <https://doi.org/10.1111/risa.12262>.
- Savelli, S., and S. Joslyn, 2013: The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Appl. Cognit. Psychol.*, **27**, 527–541, <https://doi.org/10.1002/acp.2932>.
- Schwartz, D. R., and W. C. Howell, 1985: Optional stopping performance under graphic and numeric CRT formatting. *Hum. Factors*, **27**, 433–444, <https://doi.org/10.1177/001872088502700407>.
- Simmons, K. M., and D. Sutter, 2009: False alarms, tornado warnings, and tornado casualties. *Wea. Climate Soc.*, **1**, 38–53, <https://doi.org/10.1175/2009WCAS1005.1>.
- Tak, S., and A. Toet, 2014: Color and uncertainty: It is not always black and white. *16th Eurographics Conf. on Visualization*, Swansea, United Kingdom, Eurographics Working Group on Data Visualization and IEEE Visualization and Graphics Technical Committee, 55–59, <https://doi.org/10.2312/eurovisshort.20141157>.
- University of Washington, 2019: Quick stats of student enrollment. UW Student Data, <https://studentdata.washington.edu/quick-stats/>.
- Vosgerau, J., 2010: How prevalent is wishful thinking? Misattribution of arousal causes optimism and pessimism in subjective probabilities. *J. Exp. Psychol. Gen.*, **139**, 32–48, <https://doi.org/10.1037/a0018144>.
- Weber, E. U., and D. J. Hilton, 1990: Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *J. Exp. Psychol. Hum. Percept. Perform.*, **16**, 781–789, <https://doi.org/10.1037/0096-1523.16.4.781>.
- Wogalter, M. S., A. B. Magurno, A. W. Carter, J. A. Swindell, W. J. Vigilante, and J. G. Daurity, 1995: Hazard associations of warning header components. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, **39**, 979–983, <https://doi.org/10.1177/154193129503901503>.
- , V. C. Conzola, and T. L. Smith-Jackson, 2002: Research-based guidelines for warning design and evaluation. *Appl. Ergon.*, **33**, 219–230, [https://doi.org/10.1016/S0003-6870\(02\)00009-1](https://doi.org/10.1016/S0003-6870(02)00009-1).
- Wu, H.-C., M. K. Lindell, and C. S. Prater, 2015a: Process tracing analysis of hurricane information displays. *Risk Anal.*, **35**, 2202–2220, <https://doi.org/10.1111/risa.12423>.
- , —, and —, 2015b: Strike probability judgments and protective action recommendations in a dynamic hurricane tracking task. *Nat. Hazards*, **79**, 355–380, <https://doi.org/10.1007/s11069-015-1846-z>.