The Impact of Color-Coded Uncertainty on Understanding and Decision-Making

Gala Gulacsik

A thesis in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Susan Joslyn

Chantel Prat

Program Authorized to Offer Degree:

Psychology

University of Washington

Abstract

The Impact of Color-Coded Uncertainty on Understanding and Decision-Making

Gala Gulacsik

Chair of the Supervisory Committee:

Susan Joslyn, Ph.D.

Department of Psychology

Currently, severe weather risk is communicated using “Watches” and “Warnings,” although their effectiveness is debated. Indeed, research suggests that explicit numeric probabilities improve people’s understanding of risk as well as the quality of their decisions (Joslyn & LeClerc, 2013). In many applied contexts however, color-coded risk is promoted as more accessible despite minimal evidence supporting this claim. This experimental study compared the effect of three forecast formats – watch & warning, color-coding, and percent chance – on understanding of event likelihood, trust in the forecast, and decision quality. Participants experienced forty virtual storms with the potential to produce tornadoes. For each storm, participants made a series of decisions about taking shelter. Participants made the most frequent cautious decisions in the watch & warning condition. However, decision quality, understanding and trust were highest in conditions with numeric probabilities.

Residents of tornado-prone areas in the U.S. are injured or killed every year. In the 2011 tornado outbreak, 553 people were killed despite accurate and timely forecasts with considerable lead time. The 2011 tornado season resulted in the highest number of fatalities in the past three decades (NOAA/NWS, 2019). Of note, the Joplin 2011 tornado remains the single most deadly tornado in the U.S. since 1953, taking 162 lives (Paul & Stimers, 2012). There is a growing consensus that deadly outcomes such as this may be due at least in part to the influence of psychological factors on public response to warning forecasts. Indeed, there is evidence that warning forecasts have not had the intended influence on public response. For instance, in a study looking at tornado seasons 2008 to 2010, the likelihood of taking shelter was no greater for those under a warning than for those who were not warned but living in the same county (Nagele & Trainor, 2012). While there are many reasons for non-compliance – some outside of the control of the resident, such as lack of adequate shelter – the effectiveness of the risk communication itself may be a contributing factor. This begs the question, if forecasts are accurate and timely, could it be that the forecast information is not interpreted as is intended? And is this due to how the forecast information is presented to the public?

The National Weather Service currently communicates tornado risk to the public using a watch and warning forecast. A ‘watch’ means tornadoes are possible in and near the watch area. A ‘warning’ means that a tornado is imminent and taking shelter is advised. We refer to these latter forecasts as deterministic because they imply that a tornado will definitely occur. In fact the probability of a tornado strike varies geographically within the warned area. Some locations are more likely to be hit while others are less likely. However at present this information is not made available to those in the general public.

Evidence suggests that people understand that all forecasts involve some level of uncertainty (Joslyn & Savelli, 2010), probably due to their own prior experience with similar forecasts. For instance, in tornado seasons May 2011 to May 2014, ninety-four tornadoes occurred in and around Birmingham, Alabama. For the 132 tornado warnings issued, 42% of them were hits (i.e. warning with a tornado outcome) and 58% were false alarms i.e. warning without a tornado outcome (U.S. National Weather Service, 2011). Because present day warnings do not make the uncertainty explicit, residents may regard many as "wrong" and ignore future forecasts. In other words, they may not trust them.

At least part of noncompliance with warnings may be due to lack of trust in the warnings themselves which may in turn, have to do with presenting them as deterministic. In this context, we define trust as the degree to which the information appears reliable, adequate or complete. Indeed, previous research suggests that people fail to take protective action because they do not believe that the warning information is sufficient to justify the costs associated with taking action (Nagele & Trainor, 2012). Additionally, experimental evidence suggests, that when weather outcomes differ from single value forecasts, users experience it as a false alarm and their trust in forecast is lowered (LeClerc & Joslyn, 2015). This may be because evidence shows that people intuitively know that forecasts are uncertain; expect a range of outcomes with deterministic forecasts (Joslyn & Savelli, 2010); and prefer forecasts that include uncertainty information (Morss, Demuth, & Lazo, 2008). In the long run, lowered trust may lead to ignored warning forecasts. In the same study (LeClerc & Joslyn, 2015), reduction in trust was not nearly as great in a condition that also included an uncertainty estimate. So, when a user is given a deterministic forecast, they may regard it as incomplete and therefore not trustworthy because of the lack of explicit uncertainty information. Even so, scientists are reluctant to incorporate expressions of

likelihood into forecasts for public consumption because of the perception that people struggle to interpret probabilistic information (Gigerenzer, Hertwig, Van, Fasolo, & Katsikopoulos, 2005) which some believe is compounded by low numeracy of the U.S. population (National Research Council, 2006). However, fears that the public cannot understand probabilistic information may be ill-founded. Although there is new evidence to suggest that a small group of very low numerate individuals may not benefit from numeric probability forecasts, the forecasts do not decrease decision quality compared to conventional deterministic forecasts. Moreover, all other groups benefit from probabilistic forecasts, have greater trust in the forecast, and make better decisions regardless of level of education (Grounds & Joslyn, 2018).

Mistrust in forecasts may lead to delay in taking precautionary action or ignoring the forecast all together. In the case of a tornado, forecasts are developed over time. In the days and hours leading up to a tornado strike, models use atmospheric conditions to make estimates about where tornadoes are likely to form. As the event gets closer, meteorologists determine a more precise threat area. During this time, residents may receive multiple forecast updates with the current threat level. With each update the information may change, requiring residents to replace old information with current information. Ultimately, residents must make a final decision about whether or not to seek shelter. Evidence suggests that people delay responding to weather warnings in order to seek additional information, in some cases needlessly exposing themselves to danger (e.g. Mileti & Fitzpatrick, 1991). Thus, at least part of the problem with public response to warning forecasts may be that people may continue to seek information simply because they do not regard warnings in their present form as sufficient. This is particularly important when lead-time is short, as with tornado warnings. Time for evacuation or taking shelter may be unnecessarily spent on collecting additional information. Delay behavior may be

attributed to a tendency in which people continue to gather information despite the fact that enough information is already at hand to inform a sound decision, a phenomenon referred to as delay beyond optimal stopping (Hershman & Levine, 1970). The behavior is pronounced under situations of time pressure, not unlike that experienced during severe weather warnings (Schwartz & Howell, 1985).

It may be that with the tornado hazard, if forecasts included some estimate of uncertainty, residents would regard it as more complete and trust it more. Indeed, as mentioned above, there is strong experimental evidence that people make better decisions and have greater trust in the forecast if the likelihood is made explicit (i.e. 30% chance of > 24" accumulation; Joslyn & Leclerc, 2013). Including likelihood information in forecasts may confer these benefits because it enhances transparency but also because it provides valuable information. However, the research to date has tested fairly simple single decision situations. It is unclear whether these positive effects will hold in a dynamic decision environment – a situation with changing information, as is the case with approaching tornados or hurricanes. Specifically, it is not known whether people are capable of comprehending multiple sequential probabilistic forecast updates nor whether the updates help them determine when their individual risk is high enough to justify the cost to take precautionary action.

Therefore, likelihood expressions may need to be simplified for dynamic decision environments to allow for rapid and easy understanding. Color-coding is used in many sectors to convey risk information and has been integrated into risk communication tools currently being developed by the National Oceanic and Atmospheric Administration (NOAA). However there is very little research on how users understand color coding. There is evidence that color enhances salience (Wogalter, Conzola, & Smith-Jackson, 2002) and that color warning labels are perceived

to be more hazardous than those shown in black and white overall (Braun, Kline, & Clayton Silver, 1995). Furthermore, field research shows that hurricane evacuees preferred color-coded depictions of probability to a black and white trajectory forecast (Radford, Senkbeil, & Rockman, 2013). In another study, a "traffic light" configuration (red, orange, green) was rated by participants as a suitable representation of uncertainty (Tak & Toet, 2014). Color has also been shown to correct the wrong assumption that tornadoes are more likely to occur in the center of the warned area (Ash, Schumann, & Bowser, 2014).

Color-coded risk communication is often employed in real-world settings, but not without problems. Up until 2011, the Department of Homeland Security in the United States used the Homeland Security Advisory System (HSAS). HSAS was a graduated hierarchy of terrorist threat represented by colors, words, and phrases, although the government did not test whether the public could understand the system prior to implementing. A study conducted in 2004 after the system was implemented, showed that more than half of the participants (57.8%) ranked the colors used in an order that conflicted with that employed in the actual HSAS scale (Mayhorn, Wogalter, Bell, & Shaver, 2004).

In fact, the literature shows many variations of interpretation of risk expressed with color. With the exception of red, often found to convey the notion of greatest risk (Borade, Bansod, & Gandhewar, 2008; Hellier, Tucker, Kenny, Rowntree, & Edworthy, 2010; Kline, Braun, Peterson, & Silver, 1993), there is little consensus on the rank order of colors to convey risk. For example, one study suggested that users do not differentiate yellow and orange with regard to hazard (Chapanis, 1994). In another study, yellow was shown to communicate greater hazard than orange (Wogalter et al., 1995). Blue and green are considered lower hazard than other colors, but blue was not rated as significantly different in hazardousness than green (Braun, Kline, & Silver,

1995; Rashid & Wogalter, 1997). In addition, there are cultural differences. Orange, rather than red, is considered of greatest hazard by Chinese participants (Lesch, Rau, Zhao, & Liu, 2009). The lesson here is that, just because users prefer some form of communication does not mean that they understand it (only that they think they do).

There is also a potential for people to misattribute other information types to the colors of color-coded likelihood. Often weather-related information is color-coded to display quantities such as wind speed and inches of precipitation. Because of this precedent, color-coded likelihood information may be misinterpreted as a quantity, such as the severity of the storm or the extent of the damage, rather than a likelihood. In other words, prior experience with other color-coded weather information may confuse the interpretation of color-coded likelihood information. Indeed, field interviews show that people tend to search for cues about severity in the time leading up to a tornado. Colorful radar imagery was reported as a prominent cue for severity. While it is true that color depicted rainfall quantities in this context, very few respondents could explicitly name what it was measuring and yet they associated colors with severity, nonetheless (McPherson et al., 2013). This suggests that people may make assumptions about color based on pre-existing interpretations.

It is important to note that the studies reviewed thus far, tested the rank ordering of risk conveyed by color alone. No studies, of which we are aware, have asked what numeric likelihoods participants would assign to each color category and how that compares with what was intended (e.g. “a 40% chance of showers and thunderstorms”).

Even when color is not involved, there is evidence for a tendency to misinterpret graphically depicted likelihood information as some deterministic quantity. For example, users were shown to misinterpret visualizations of percent chance of precipitation as duration or

geographic extent of precipitation (Joslyn, Nadav-Greenberg, & Nichols, 2009). In addition, previous research showed that visualizations with a range of possible outcomes (e.g. 3-5" precipitation) were thought to represent diurnal fluctuation – that is, they interpreted it as a deterministic forecasts with a nighttime low and a daytime high (Savelli & Joslyn, 2013). It may be that people have a general tendency to misinterpret likelihood expressions as some deterministic quantity if the expression permits it, suggesting a "desire for certainty" (Slovic & Lichtenstein, 1971). This may have to do in part with cognitive load. A probabilistic forecast communicates that multiple outcomes are possible and thus, demands a larger cognitive load to comprehend than a forecast with a single outcome. It is possible that users avoid the more difficult interpretation and instead choose the easier one via "attribute substitution" (Kahneman & Frederick, 2002). Evidence suggests when probabilistic information is interpreted in this way in particular, it leads to systematically different decisions that may have dangerous consequences (Savelli & Joslyn, 2013).

In sum, public trust in the forecast may be critical for good and timely decisions about precautionary action. Trust may be maintained by providing understandable likelihood information. The question remains how best to convey likelihood information in a dynamic decision environment to improve public response to warning. Can users utilize numeric likelihood information in this environment or is the cognitive load too great?

In reality, assigning a single probability may not be possible. In practice, a forecast may be expressed as a second-order uncertainty (i.e. a range of percent chance) - situations in which there is uncertainty about the uncertainty; 10 – 20% chance of light rain, for instance. Forecasters may opt to communicate such ambiguity because it conveys possible outcomes as well as acknowledges the limitations of the underlying model. If users can utilize numeric likelihood

information in a dynamic environment, can they also make good use of second-order uncertainty information? If they cannot, can color-coding communicate the same precision of likelihood information without the detrimental effects?

The study reported here investigated these questions by comparing the effect of forecast format on decision quality, timeliness of the decision, trust in forecast, and understanding via a computerized decision task similar to that of the experiment by Schwartz and Howell (1985). We tested four experimental formats of forecast uncertainty: color-coded, color-coded with a range of percent chance (color + percent chance range), range of percent chance (percent chance range), and a single percent chance value (percent chance). A watch & warning forecast format was used as a control condition to emulate real-world forecasts.

Method

Participants

Of the 489 University of Washington students who participated in the study, 57% were female, 42% were male, and 1% undisclosed. All participants were between 18 and 24 years of age with an average age of 19 years and received extra credit for participation.

Apparatus

The tornado simulation was programmed using an online html platform. The simulation was administered on standard desktop computers.

Procedure

Approximately ten participants participated in each one hour session. After providing informed consent participants spent about 30 minutes on the decision making task (Mean = 30.15, Min = 18.64, Max = 50.67) excluding the instruction phase which took approximately 15 minutes (see Appendix A for instructions). The researcher read the instructions aloud as

participants read along on the computer. The instructions provided background on the tornado hazard including: how a tornado is formed; the wind speeds of weak (73-112 miles per hour (mph)) and strong (260 + mph) tornadoes. They explained that strong tornado “can level and blow away almost any house and its occupants” and that even weak tornadoes can cause injury due to flying debris. In addition, participants were told that although tornadoes occur across the country, they are most common in the central United States in an area called "Tornado Alley.” Severity was held constant by telling participants that all storms produce wind speeds at 90-112mph, consistent with weak tornadoes. Participants were told they would experience 40 storms with the potential to produce tornadoes. Every storm moved west to east, from the same distance, towards “home,” a house in which the participant was to imagine they were located. For each storm, there were seven decision points at which the storm moved closer to home. The instructions described the decision options as: wait for more information; take shelter in a tornado shelter near your house; or not take shelter. The shelter location was described as being outside of the house (as opposed to within) to emphasize the increasing danger of taking shelter as time passes and the storm comes closer to their home. A decision to take shelter or not take shelter was a final decision for that round. After making a final decision, the participant saw the rest of the information for that round, answered a few more questions, and learned whether or not a tornado hit home.

The instructions then described the point structure, which was intended to simulate the cost loss structure of real-world decisions and to incentivize participants’ performance. Participants had 24,000 points to start. The starting budget was tailored to a participant using the costliest strategy to ensure participants did not run out of points before the end of 40 trials. The goal was to complete 40 trials with the highest possible point budget remaining. Participants

were told there was no cost to wait on decision points 1-3; on decision points 4-7 there was a 20-point cost for every wait decision to reflect the increasing danger of the storm approaching home. If a tornado hit home and the participant had chosen to wait or not take shelter as their final decision, their point balance was reduced by 1500 points. The cost to take shelter began at 303 points at decision point 1 and increased per decision "because taking shelter can be costly in terms of effort, time, and sometimes money… [and] to reflect the increasing danger of being caught in a vulnerable position when a tornado strikes."

$$Shelter\ Cost = 300 + 3 * decision\ point^2 \quad (1)$$

If participants took shelter, they paid a one-time cost that guaranteed safety, and thus no further loss was possible on that trial. All costs and penalties were deducted immediately from the point balance, which was shown on the screen at all times. The cost of decisions is shown in Table 1.

Table 1

Cost of Decisions

Decision	Cost
Wait	Decision points 1-3: no cost Decision points 4-7: 20-points per wait decision
Take Shelter	$Shelter\ Cost = 300 + [3 * decision\ point^2]$
Not Take Shelter	No cost

Participants were informed that the experiment began with 10 practice trials followed by 40 data collection trials. Participants were told that each trial represented a storm independent from the others to discourage them from deducing trends in the weather conditions.

The decision task consisted of a series of storms each with a unique path. The storm moved across a virtual grid (shown in Figure 1) toward the participant's home at the far eastern

boundary, although this graphic was not shown to participants. For explanatory purposes, each cell on the grid is numbered and increases North to South (latitudes 1-7) and West to East (longitudes 1-8). The storm movement was computer generated in real-time based on the probability by which the storm moved forward to a cell in the adjacent longitude. At the start of every trial, the storm path began at latitude 4, longitude 1. The location of home was the same for every trial (latitude 4, longitude 8). At each decision point, the storm advanced to a new cell, one longitude closer to home. When the storm was located in a cell within latitudes 2-6, the storm could advance to one of three cells in the adjacent longitude. There was a 0.3 probability of moving to the cell Northeast or Southeast, or a 0.4 probability of moving laterally to a cell to the East. When the storm was located along the edges of the grid (cells within latitude 1 or 7) it could advance to one of two cells. There was a 0.7 probability of moving laterally to the East, or a 0.3 probability of moving to a cell towards the center (i.e. Southeast from latitude 1 or Northeast from latitude 7). Figure 2 shows the probability of the storm being in each cell.

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

Figure 1. Example of storm movement across grid.

Latitude \ Longitude	1	2	3	4	5	6	7	8
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

Figure 2. Probability of Storm being in each Location

Participants received seven weather forecasts over the course of a trial, one for each decision point, each forecast corresponded to the storm's movement to a new cell at a successive time point. Upon receiving a forecast, participants were asked to rate the likelihood of a tornado hitting their home by clicking on a two-inch, unmarked line centered on the screen and labeled at its left end with "impossible" and the right end with "certain." Participants were asked to rate the damage they expect if a tornado were to hit the house on an identical line just below labeled to its left and right "not severe" to "very severe," respectively. On the next screen, the same forecast was displayed again; participants were asked what they would like to do and chose by clicking on one of three radio buttons : wait ([cost in points]); not take shelter (0 points if not hit; 1500 points if the tornado hits); take shelter ([cost in points]). Participants were asked to rate how much they trusted the forecast information by clicking on an unmarked line, anchored on the left with "not at all" and to the right with "completely." If the participant chose to wait, they moved to the next decision point, received an updated forecast, and were asked to answer the same questions as outlined previously. If they made a final decision to "not take shelter" or "take shelter", in order to discourage rushing through trials by deciding early, they saw the remaining forecast updates and were asked to answer the same questions regarding likelihood, severity, and trust although they were not allowed to change their decision. In the subsequent sections, forecast updates will be referred to as *decision points* irrespective of whether a final decision has already been made for that trial. After all seven decision points, the outcome screen gave a summary of the trial in the following format: "The tornado [missed your area / hit your home]. Your final choice was to [take shelter / not take shelter/wait]. Your actions cost you [points spent]. Taking shelter was [not] necessary. You [avoided a 1500-point penalty / were penalized 1500 points]." The next page asked again how much they trusted the information they were

given. After clicking next, the screen displayed a congratulatory message with the ending point balance, the number of times a tornado hit home, the number of times the participant chose take shelter in these instances, and the penalties avoided. The message also listed how many times the tornado missed home, and the number of times of these instances the participant chose take shelter in when it was not necessary.

Table 2 lists all questions, their frequency and timing, the format in which the participants responded, and the accompanying information displayed with the question. Questions grouped by a bold border were shown together.

Table 2

Experimental Questions, Answer Format, Timing, and Accompanying Information

		Answer Format		Information Displayed	
Question Type	When Shown	Radio Button	Slider	Forecast	Point Budget
Likelihood Rating	Each forecast update/decision point (7x per trial)		x	x	x
Severity Rating	Each forecast update/decision point (7x per trial); and 1x end of trial		x	x	x
Decision (i.e. wait, take shelter, not take shelter)	Each forecast update prior to and including final decision (up to 7x per trial)	x		x	x
Trust Rating	Each forecast update/decision point (7x per trial); and 1x end of trial		x	x	x

After completing the experiment, participants were provided a debriefing sheet (Appendix B). The debriefing sheet explained that the intent of the experiment was to evaluate the effect of display format with likelihood information on human performance. Next, participants were directed to complete a web-based debriefing survey (Appendix C). The survey asked open-ended questions about the participants' approach to making decisions and what

information was used in making decisions. Additionally, it asked whether they used a strategy in completing the simulation and to describe it. The survey also requested they describe how they interpreted the information about the storms. Finally, it asked them to list the information they used to estimate severity of damage.

Participants were given course credit and rewarded in cash commensurate to their performance and the rate of pay. Participants received payment if they ended the experiment with more than 11,880 points. We determined the threshold for payment by calculating the cumulative cost if participants chose take shelter on the first decision point of every trial: 303 points multiplied by 40 trials (12,120 points). If participants were to make these decisions, they would have 11,880 points remaining at the most. This payout threshold was intended to encourage participants to engage with the task as opposed to taking the easy way out to complete it quickly. Participants earned $1 for every 1500 points above the 11,880 point threshold.

Stimuli

All participants received the same background information on the tornado hazard and instructions on the decision task. All forecasts were derived from the probability of hitting home from the cell in which the storm was hypothetically located (Figure 3). The probabilities of a tornado hitting home were realistic and ranged from 0 to .40. After each forecast update, the computer generated the next storm movement to an adjacent cell and displayed the forecast relevant to the probability of hitting home from that cell.

Latitude \ Longitude	1	2	3	4	5	6	7	8
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

Figure 3. Probability of tornado hitting home from each cell on the grid.

In this experiment, there were five conditions that varied the format by which the same underlying information about tornado probability is communicated to participants. A watch & warning forecast format was the control condition as it represented how tornado forecasts were presently communicated. The four experimental conditions expressed tornado probability with color-coding, percent chance, ranges of percent chance, and a combination of color-coding and percent chance ranges.

In the watch & warning condition, participants received a message indicating whether the area was under a watch, warning, or neither. Definitions for each, taken from the National Weather Service, were provided to participants in the instructions. Participants were told a ‘watch’ meant that tornadoes were possible in or near the watch area. A watch forecast was shown if the tornado reached a position at which the probability of hitting home was greater than or equal to 0.13 and less than or equal to 0.24. In this range of probability, it was the

economically optimal decision to wait. The optimal decision was determined by selecting the decision in which the participant could expect to lose the least points, or the smallest expected loss. Expected loss was based upon the probability of a tornado hitting home at present and the cost and loss of decision. For positions in which the probability of hitting home was less than 0.13, no watch or warning was issued. In this range of probability, the appropriate decision was to not take shelter. A warning was issued at positions at which the probability of hitting home was .25 or greater. In this range of probability, the appropriate decision was to take shelter. Participants were told a 'warning' meant that a tornado had been sighted or indicated nearby and may enter the warning area. Table 3 shows the stimuli and the optimal decision for the given range of probability. The optimal decisions for each range held true for the experimental conditions' stimuli as well (shown in Table 4).

Table 3

Range of Percent Chance and Optimal Decision

Percent chance	Optimal Decision
$< 13\%$	Not shelter
$\geq 13\%$ and $\leq 24\%$	Wait
$> 24\%$	Take Shelter

Table 4

Forecast Format (identified by A-E) and Stimuli. Preceded by "Chance of tornado hitting your area:"

A: Watch & Warning	B: Color	C: Percent Chance Range	D: Percent	E: Color + Percent Chance Range
No watch or warning		$0 - 12\%$	6%	Stimulus B with C
Watch		$13 - 24\%$	19%	Stimulus B with C
Warning		$25 - 40\%$	33%	Stimulus B with C

Participants in the color condition received a color-coded forecast described as the "Chance of tornado hitting your area." They were told that warmer colors indicated higher likelihood of a tornado hitting their area. This was intended to emulate the Forecasting a Continuum of Environmental Threats (FACETs) framework, a proposed next-generation severe weather watch and warning framework administered by the National Weather Service. Green, yellow and orange were displayed individually indicating the present level of likelihood (see Figure 4). A green bar was shown when the probability was less than 0.13; yellow ≥0.13 and ≤0.24; orange >0.24 and was based on the probability of a tornado hitting home from the cell in which the storm was presently positioned. Table 4 shows the forecast stimuli in watch & warning, color, percent chance ranges, percent chance, and color + percent chance ranges. Notice that these values are identical to those that define "no watch or warning," "watch" and "warning" in the condition above. Stimuli in the percent chance condition were drawn from a continuous scale from 0 to 40%. Each example (shown in column D) was drawn from the three percent chance ranges in column C.

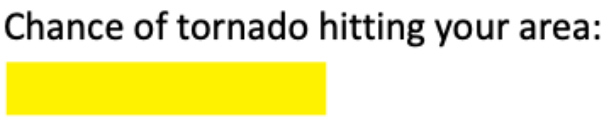

Figure 3. Forecast stimuli in color condition.

While the watch & warning and color conditions were categorical, the percent chance condition displayed likelihood on a continuous scale from 0-40% representing the actual probability of the tornado hitting home from that cell rounded to two decimal places and presented as a percent chance. The difference between the percent chance and percent chance ranges condition was that those in the percent chance condition saw a single value forecast that lied within a given range, while those in the percent chance range condition saw the range as

shown in Table 4. Participants in the color + percent chance range condition saw both the color bar and the analogous percent chance range. Each forecast stimulus in the experimental conditions was preceded with the following text: “Chance of tornado in your area:” (e.g. Chance of tornado in your area: 6%).

Figure 5 shows the grid across which the storm moved and the stimulus that was shown at each position. Optimal decisions are shown in shades of blue. It is important to note that in conditions with categorized expressions of likelihood, their category boundaries were set at the optimal likelihoods for each of the three decision choices.

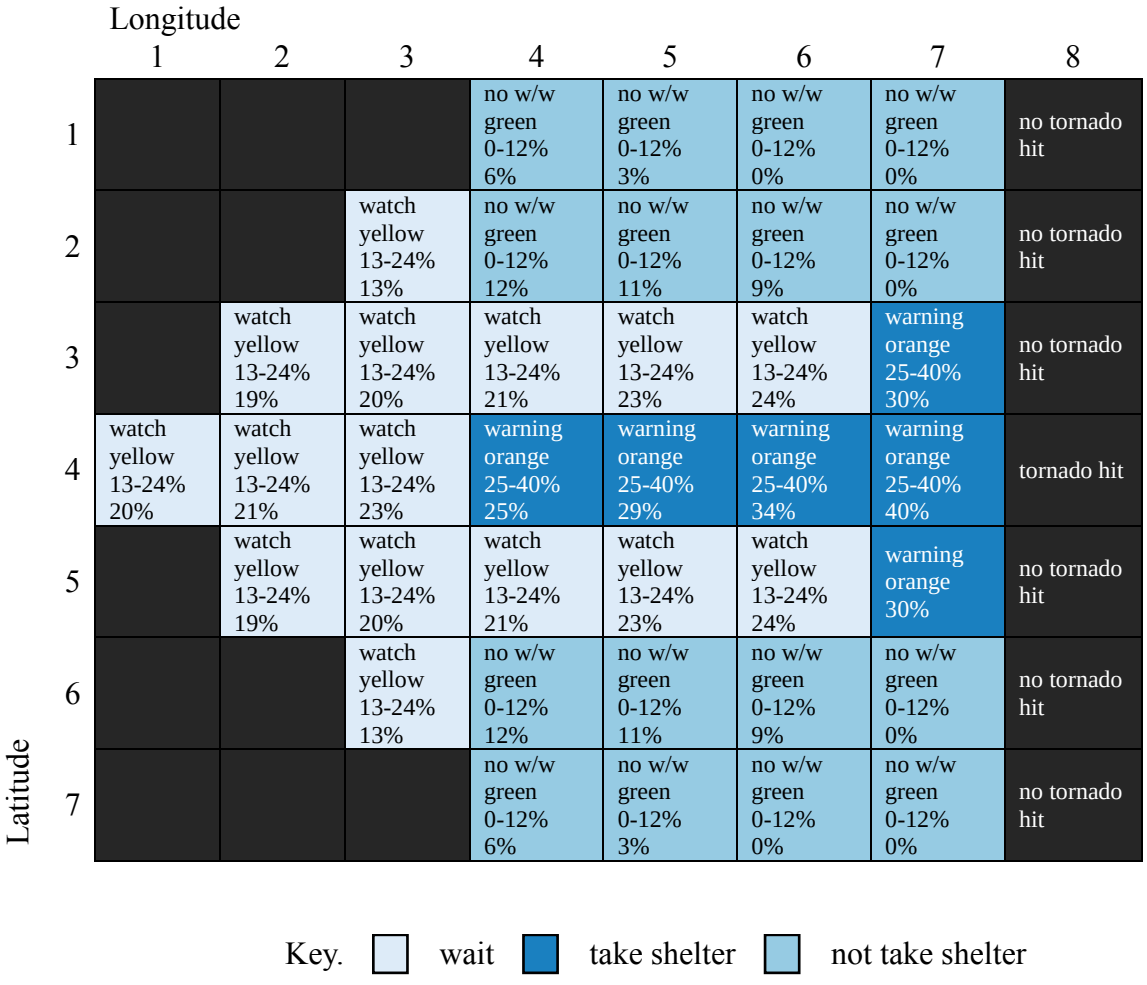

Figure 5. Storm location grid with stimuli of each forecast format and optimal decision.

Design

This experiment was a one-factor independent groups design with five between groups levels. Participants were randomly assigned to one of five forecast formats, or levels: watch & warning, color, percent chance, percent chance range, and color + percent chance range. There were four dependent variables: forecast understanding, decision quality, decision timeliness, and trust in the forecast. Forecast understanding was operationalized in two ways. The first, addressed the similarity between participants likelihood ratings and the expression of probability in the forecast – referred to as *actual percent chance* in subsequent sections. The second, evaluates whether forecasted probability information was mistaken for indications of severity and how that influenced participant severity ratings. Decision quality was represented by the mean expected value of the participants' decisions. Decision timeliness was represented by comparing the point at which a participant made a final decision to the optimal decision point, determined by a comparison of the expected value of each of three alternatives at each time point. Finally, trust in the forecast was determined by examining participants' trust ratings.

Results

The results are presented here in two parts. The first section describes the Analyses of Variance ANOVAs testing the effect of forecast format (watch & warning, color, color + percent chance range, percent chance range, and percent chance) on understanding (event likelihood and severity) and trust in forecast information. The second part describes the ANOVAs on of forecast format on decision quality and decision timeliness. Planned contrasts were corrected for familywise error using the Bonferroni correction (α=.0125). Post hoc comparisons were corrected for multiple comparisons using the Tukey test.

Part 1: Understanding of and Trust in the Forecast

To evaluate the effectiveness of communicating likelihood information with the five formats, understanding was operationalized as the difference between the participants' likelihood ratings and the actual percent chance of the tornado hitting home from the storm position at that decision point. Recall that participants made likelihood ratings by moving a handle on an unmarked line anchored on the left by "impossible" and on the right by "certain". Responses were translated to the percentage of the line between the left anchor and the handle. For a given storm path, the probability of a tornado hitting home (expressed 0 to 1, to which we refer as the "actual percent chance") at each position on the grid was subtracted from the participant's likelihood rating at that position (i.e. likelihood rating – actual likelihood). A mean likelihood difference score was calculated by summing all differences for all forecast updates and dividing by the total number of forecast updates over forty trials. Negative numbers meant participants' likelihood ratings were lower than the actual percent chance of a tornado hitting home while positive numbers meant it was greater. Negative and positive signs were maintained to determine whether there was a general bias in the estimate.

In order to understand whether accuracy of likelihood estimation differed by format or range of actual percent chance, we compared accuracy by forecast format and the three actual percent chance ranges, (e.g. green + 0-12%, yellow + 13-24%, orange + 25-40%; 0-12%, 13-24%, 25-40%) referred to as "actual percent chance ranges" hereafter (see Figure 7 and Table 5).

Table 5

Differences of Likelihood Rating and Actual Percent chance

Forecast Format	Likelihood Difference				
	Actual Percent chance range			Overall	
	0-12%	13-24%	25-40%	(0-40%)	One sample t-test (value = 0)
Watch & Warning	*M=11.20 SD=12.1*	*M=23.3 SD=15.2*	*M=34.3 SD=17.4*	*M=21.93 SD=12.7*	t(95) = 16.92, p<.0001
Color	*M=9.76 SD=12.3*	*M=21.2 SD=13.8*	*M=42.0 SD=19.2*	*M=21.89 SD=11.95*	t(100) = 18.41, p<.0001
Color + Percent Chance Range	*M=5.76 SD=10.4*	*M=13.9 SD=11.4*	*M=33.7 SD=16.2*	*M=15.23 SD=10.29*	t(97) = 14.66, p<.0001
Percent Chance Range	*M=0.91 SD=10.3*	*M=7.43 SD=10.7*	*M=25.3 SD=14.6*	*M=8.89 SD=9.8*	t(100) = 9.12, p<.0001
Percent Chance	*M=-4.41 SD=7.7*	*M=6.53 SD=10.3*	*M=17.9 SD=12.9*	*M=5.69 SD=9.1*	t(92) = 6.02, p<.0001

On average participants overestimated in all forecast formats but did so to the greatest degree with watch & warning forecasts and forecasts that included color. In addition, likelihood differences were greater (positive) as the actual percent chance increased for all formats. To measure how likelihood ratings compared within actual percent chance ranges and between forecast formats, a two-factor mixed design ANOVA was conducted on mean likelihood difference scores with actual percent chance range as the within-subjects factor and forecast format as the between-subjects factor. On average, participants in all conditions showed a tendency to overestimate likelihood by 15% (*SD=12.70%*) as compared with the actual percent chance of a tornado hitting home. In fact, all likelihood differences were significantly greater than zero (see Table 5 for test statistics). There was a significant main effect of forecast format on likelihood difference, $F(4,1449) = 93.68$, $p<.001$. Overestimation was most pronounced in the color *(M=21.89, SD= 11.95)* and watch & warning conditions *(M=21.93, SD=12.70)* and least in

the percent chance condition *(M=5.69, SD=9.10)*. Planned contrasts revealed that the mean likelihood difference in the watch & warning condition (*M=21.93, SD=12.7*) was significantly greater than all of the conditions that included numeric percent chance: color + percent chance range *(M=15.23,SD=10.29),* F(1,1449)=18.42, p<.001; percent chance range *(M=8.89, SD=9.8)*, F(1,1449)=70.89, p<.001; and percent chance (M=5.69, SD=9.1), F(1,1449)=105.64, p<.001. In addition, post hoc contrasts showed that likelihood differences of watch & warning and color were not significantly different from one another, p>.0125, suggesting color and watch & warning forecasts are equally detrimental to understanding. Furthermore, post hoc comparisons showed that likelihood differences in the color + percent chance range condition was significantly greater than in the percent chance range condition, p<.001, suggesting that adding color actually made understanding worse even though the numbers were included. However, there was no significant difference between percent chance range and percent chance, p>.05, suggesting that ranges confer similar advantage to individual probabilities. There was also a significant main effect of actual percent chance range on likelihood difference scores, F(1.54, 750.54)=888, p<.0001. The degrees of freedom were corrected for violation of sphericity using the Greenhouse-Geisser correction. Overestimation was smallest in the lowest range (0-.12; *M=4.72, SD=12.09*), higher in the mid-range (.13-.24; *M=14.52, SD=14.15*), and greatest in the highest range (.25-.40; *M=30.78, SD=18.18*). See Figure 8. Although, in all conditions, likelihood difference increased as the actual percent chance increased there was a significant interaction between actual percent chance range and forecast format, F(8,1449)=4.09, p<.000. As shown in Figure 8, likelihood difference for watch & warning and color are similar in the bottom two percent chance ranges, but the overestimation for color is greater in the upper range for color *(M=42.02, SD=19.23)*.

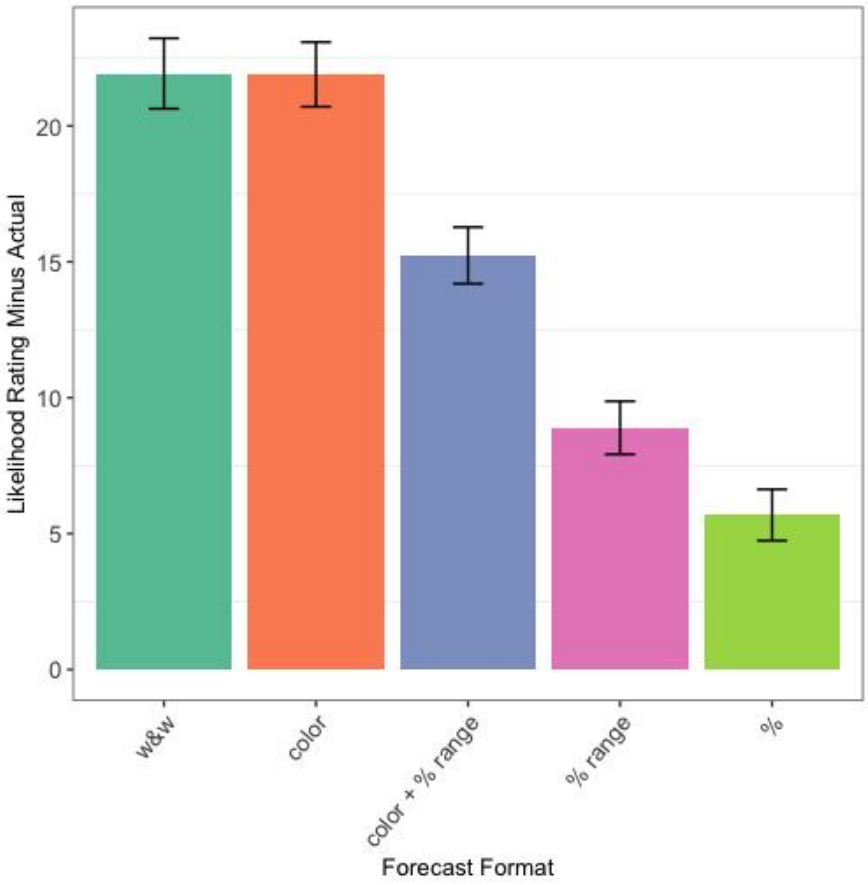

Figure 6. Mean likelihood difference by forecast format.

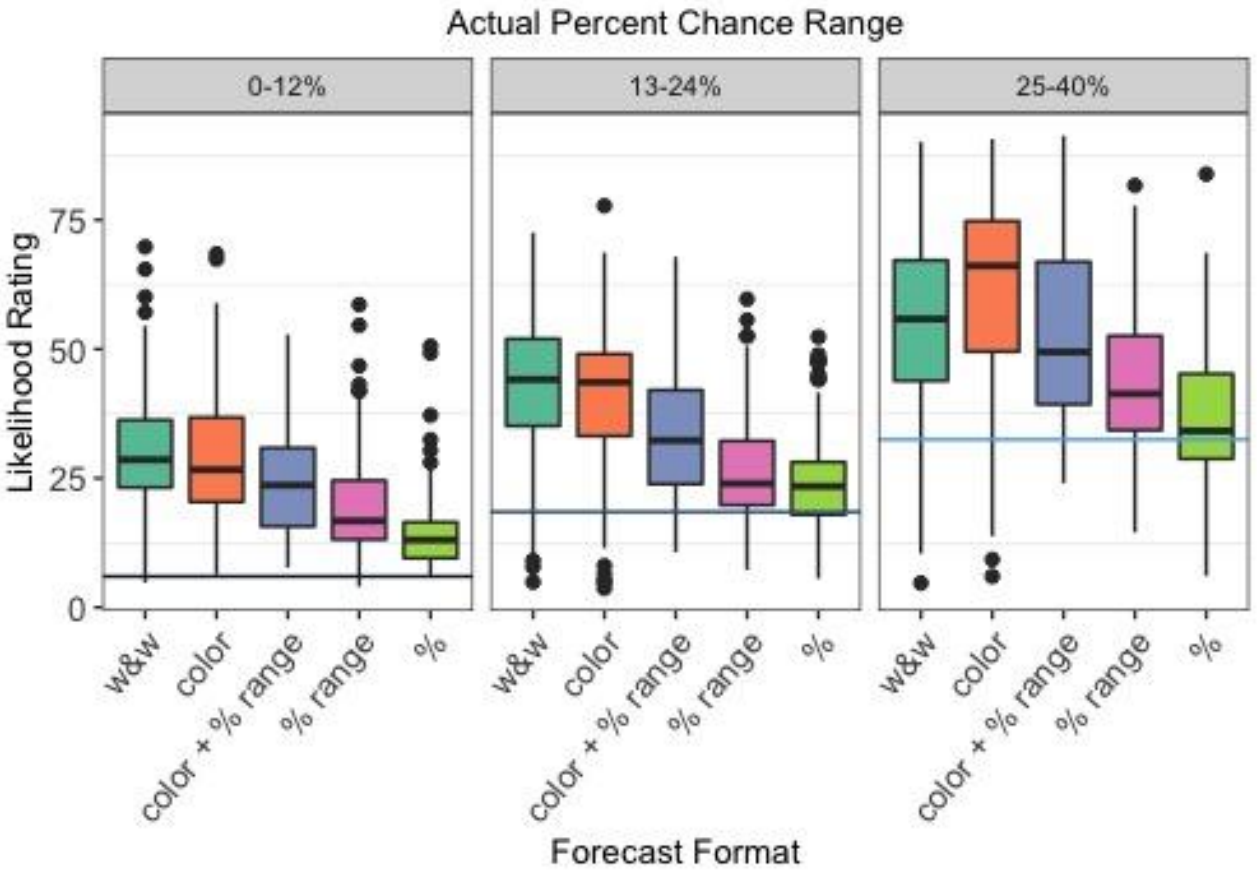

Figure 7. Box plot of likelihood ratings within categorical ranges of percent chance shown by forecast format. The black horizontal lines within boxes indicate the mean likelihood rating within forecast format for the specified actual percent chance range. The blue-scale horizontal lines that spans each actual percent chance range plot mark actual percent chance range mid-point.

In sum, participants overestimated likelihood overall, however the likelihood differences were largest in the watch & warning forecasts and those that included color. Including percent chance information with color-coded uncertainty improved the accuracy of likelihood ratings however not to the degree of percent chance alone. Surprisingly, percent chance range performed comparably to percent chance alone, suggesting understanding was substantially better not only with numeric uncertainty but also with second-order numeric uncertainty information.

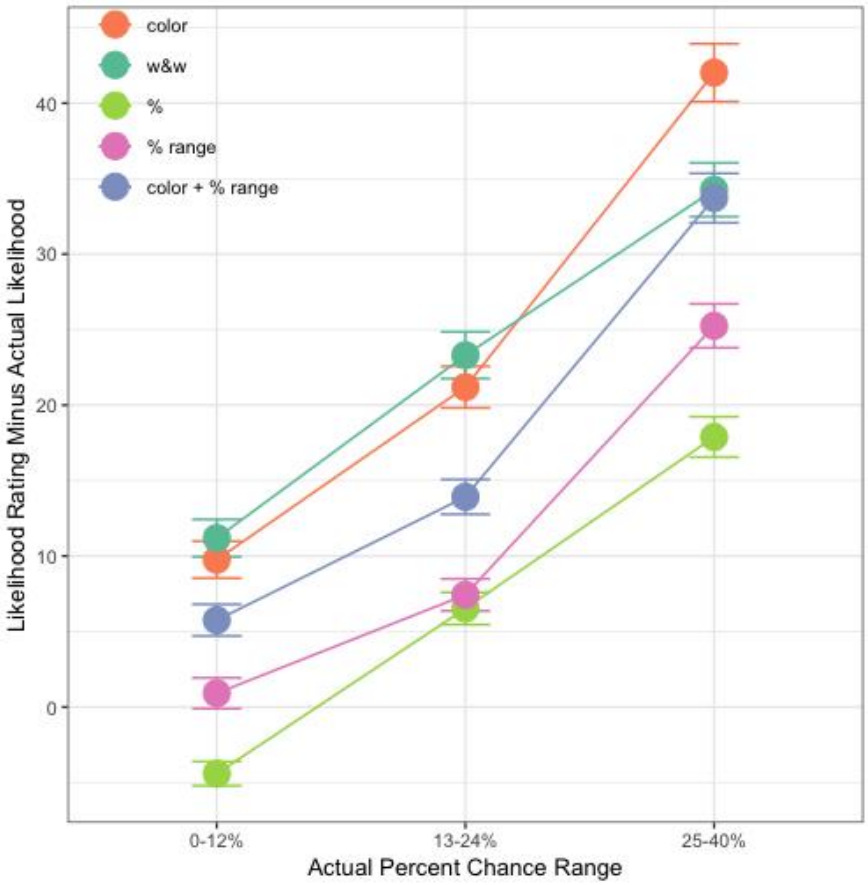

Figure 8. Average likelihood difference by Actual percent chance range and forecast format.

Another possible misunderstanding about which we were concerned, in particular with respect to color-coding, was that the expression contained information about severity as well. In addition to rating likelihood, participants made severity ratings at each decision point by moving a handle on an unmarked line anchored on the left by "not at all" (translated to 0) and on the right by "completely" (translated to 1). Responses were coded as the percentage of the line between the left anchor and the handle. 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 analyses.

For the first operationalization, the likelihood rating (expressed 0 to 1) was subtracted from the severity rating (expressed 0 to 1) at the same decision point preserving the signs (severity – likelihood) and an average was calculated for each participant. The smaller the difference (positive or negative) the greater the confusion. A positive number indicated the participants' severity rating was higher than their likelihood rating.

The difference in likelihood and severity ratings was smallest in the color and watch & warning conditions suggesting that participants interpreted them as saying something about the severity of an event (see Figure 10). A one-factor ANOVA was conducted on the difference in severity and likelihood ratings (referred to as severity difference hereafter) with the independent variable forecast format. There was a main effect for forecast format, $F(4,484)=8.05$, $p < .001$, $\eta^2=.06$. The color *(M=2.43, SD=10.45)* and watch & warning *(M=3.06, SD=15.88)* conditions showed the smallest likelihood-severity differences. Percent chance range *(M=12.92, SD=22.46)* and percent chance *(M=13.14, SD=20.84)* showed the largest likelihood-severity differences and color + percent chance range fell somewhere in the middle (*M=7.76, SD=17.75*). Planned contrasts revealed that the mean difference for color was significantly smaller than percent chance range (*M=12.92, SD=22.46*; $F(1,484)=17.25$, $p<.0001$) and percent chance conditions (*M=13.14, SD=20.84*; $F(1,484)=17.26$, $p<.0001$). This suggests that conditions that include only color give rise to the greatest confusion between severity and likelihood.

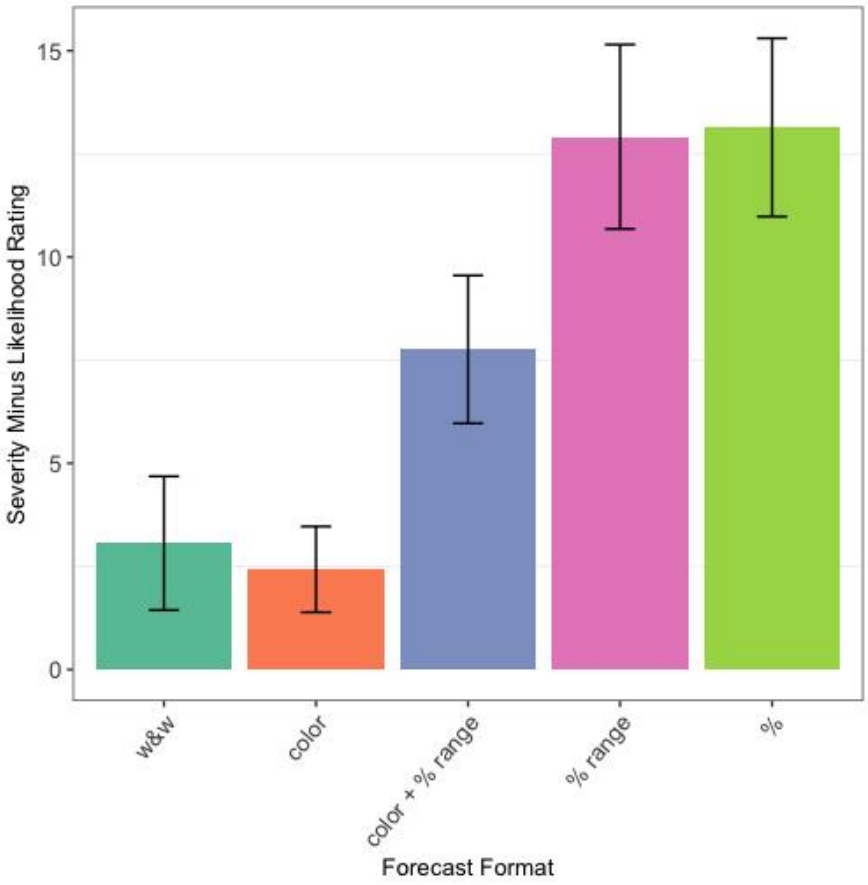

Figure 9. Bar chart of mean difference in severity and likelihood ratings by forecast format.

Mistaking likelihood for severity was also operationalized as the correlation between severity and likelihood ratings. Correlations were highest in the color condition meaning that likelihood and severity ratings were strongly associated. Correlations for each forecast format were found to be significant. Because the sampling distribution for highly correlated variables is highly skewed, the distributions were transformed into using a Fisher's "z transformation" before conducting the ANOVA. A one-factor ANOVA, with the independent variable forecast format revealed a significant main effect of forecast format $F(4,484)=6.36$, $p<.001$, $\eta^2=.05$. Planned contrasts showed that the correlation between severity and likelihood in the color condition *(M=0.79, SD=0.30)* was significantly higher than the percent chance range (*M=0.56, SD=0.43*; $F(1,484) = 20.26$, $p<.0125$), and percent chance conditions (*M=0.58, SD=0.46*; $F(1,484) =$

16.62, p<.0125). Color was not significantly different from watch & warning *(M=0.69, SD=0.36)* nor from color + percent chance range *(M=0.68, SD=0.37)* , p>.0125. This again suggests that conditions that included color gave rise to the greatest confusion. Thus, the first two operations of the severity-likelihood confusion (severity difference and severity-likelihood correlation) suggest that when numbers alone are used the misunderstanding is the least.

Third, severity ratings in the color condition showed the greatest variability, more so than all other conditions, suggesting that participants mistakenly thought the forecasts were saying something about severity and thus altering their severity rating accordingly. In order to determine whether participants inferred severity information from forecasted expressions of likelihood in a given condition, the standard deviation (SD) of each participant's severity ratings over the course of 40 trials was averaged to give a mean standard deviation by participant. At the beginning of the experiment, the instructions stated that all storms were of the same wind speed and thus the same severity. If participants understood the instructions, their severity ratings should remain constant throughout the experiment. Smaller mean standard deviation values indicated less variability and implied that participants understood that the severity of the storms remained constant and that forecasts were not communicating severity information. A one-way ANOVA on severity SD with the independent variable forecast format revealed a significant main effect of forecast format, $F(4,484)=9.72$, $p<.0001$, $\eta^2=0.07$. In planned contrasts, the severity SD of the color condition (M=18.46, SD=7.07) was significantly greater than all other conditions: watch & warning *(M=15.44, SD=6.56)*, p<.0125; color + percent chance range *(M=15.96, SD=6.87)*, p<.0125; percent chance range *(M=13.46, SD=6.08)*, p<.0001; and percent chance conditions *(M=13.46, SD=6.52)*, p<.0001. The large fluctuations in severity ratings shown in the color

condition suggest participants were misinterpreting color-coded expressions of likelihood as saying something about severity.

The three operationalizations of mistaking likelihood for severity (severity difference, severity-likelihood correlation, severity standard deviation) showed that participants indeed confused color-coded likelihood information as also saying something about the weather event's severity.

Next we examined trust in the forecast. Participants gave trust ratings on a slider anchored to the left with "not at all" (translated to 0%) and to the right with "completely" (translated to 100%). Trust ratings were given after learning the outcome of a storm at the end of each trial. Higher numbers indicate higher trust in the forecast information.

Post-outcome trust ratings were highest in the percent chance condition and lowest in the color condition. A one-factor ANOVA was conducted on mean post-outcome trust ratings with the independent variable forecast format. There was a main effect of forecast format on post-outcome trust, $F(4,484)=3.07$, $p=.016$, $\eta^2=.025$. Participants gave the lowest post-outcome trust ratings in color condition (*M=42.82, SD=15.75*) and the greatest in percent chance (*M=50.38, SD=18.80*) as well as the color + percent chance condition (*M=50.07, SD=19.0*). A post-hoc analysis showed that trust ratings for percent chance was significantly greater than color, $F(1,484)=7.56$, $p=.03$. No other significant relationships were found.

In sum, contrary to our prediction, there were no significant differences between post-outcome trust for watch & warning forecasts and the other four experimental formats. The greatest difference was seen with percent chance showing the greatest trust and color the least. Recall the color also had the highest overestimation of likelihood, which may explain the low trust. We discuss this relationship in the discussion below..

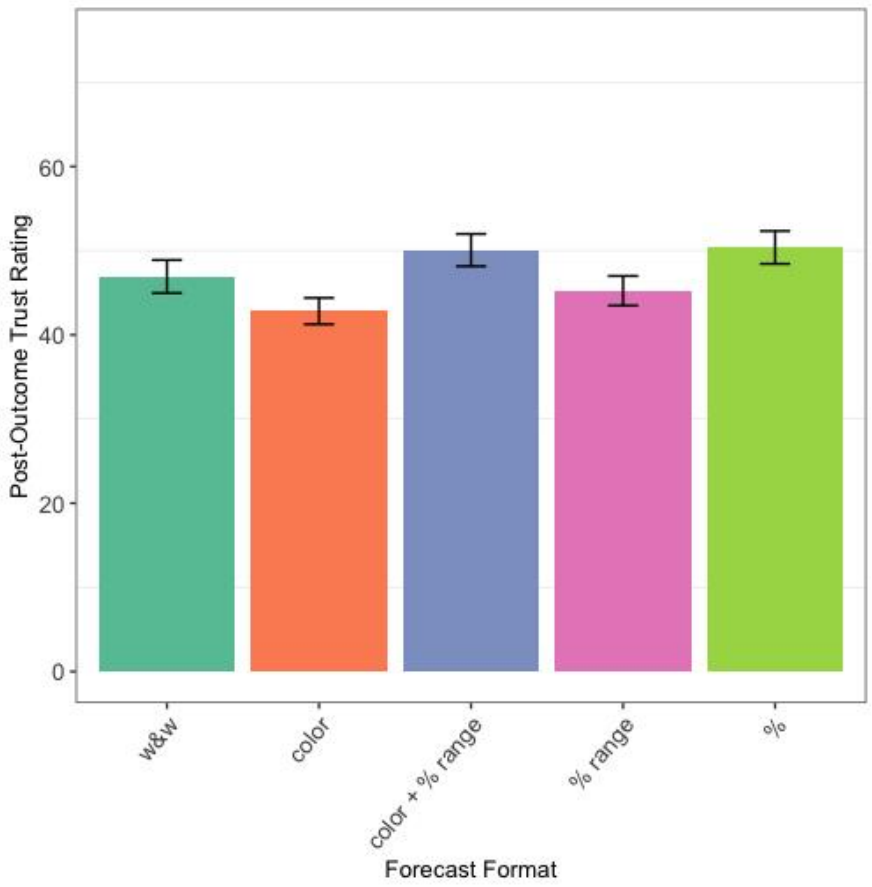

Figure 10. Bar chart of post-outcome trust in forecast by format.

Part 2: Decision Quality and Timeliness

In order to determine the impact of forecast format on decision making, decision quality was operationalized as the expected value of participants' decisions subtracted from the optimal expected value, referred to here as "expected loss" (EL) as no gains are possible. The lower the expected loss difference the better the decision. Remember that participants could make one of three decisions: *wait, take shelter, or not take shelter* If the participant chose to shelter there was a one-time cost which was the value that was assigned for that decision. The cost of waiting on the steps prior to the step on which the shelter decision was made was included. In addition, the cost to *shelter* increased as the storm moved closer to home. The cost was calculated as follows:

$$cost_{shelter} = 300 + 3 * LONG^2 + cost_{wait} \quad (1)$$

Wait or *not take shelter* decisions had an expected loss that depended on the actual probability at a given storm position across the grid. The expected loss to *not take shelter* ($EL_{not\ shelter}$) was the 1500-point hit penalty multiplied by the probability of experiencing that penalty (i.e. a tornado hit home and the participant did not take shelter). As with the $EL_{shelter}$ calculation, the wait cost is included in this decision quality calculation. The expected loss of *not take shelter* is calculated with the following formula:

$$EL_{not\ shelter} = Probability\ of\ tornado\ hitting\ home * 1500 + \ cost_{wait} \quad (2)$$

Although a *wait* decision incurred a cost under certain circumstances, like *take shelter*, the participant also assumed a risk of a hit penalty. Therefore, if "wait" was the final decision an expected loss was added to the cost. The expected loss of a *wait* decision (EL_{wait}) was first, the calculated for all cells in longitude 7 then the expected loss for the remaining cells in longitude 1-6 was backwards calculated from longitude 7. Because the participant would have chosen to wait in decision points 4-7 the 80 point prior wait cost (20 points per wait decision after longitude 3) was included. The expected loss of *wait* for longitude 7 in this table is calculated as follows:

$$\text{Longitude 7: } EL_{wait} = 80\ points\ incurred\ cost_{wait} + (probability\ of\ a\ tornado\ hitting\ home * 1500) \quad (3)$$

The expected loss of *wait* at longitude 6 and all previous longitudes was the sum of the smallest expected loss (EL_{min}) – also referred to as the optimal decision – in the adjacent East three cells multiplied by the probability of the tornado moving to the cell. Because the

probability of the storm moving to that cell varies by latitude, there are expected loss formulas for cells in the outside latitudes that differ from the calculation for those within. See the procedure in the Method section for an explanation of the probabilities that guide storm movement. The expected loss of *wait* is calculated as follows by latitude:

Latitude 1: $EL_{wait}(LAT, LONG) = 0.7 * EL_{min}(LAT, LONG + 1) + 0.3 * EL_{min}(LAT + 1, LONG + 1)$ (4)

Latitude 2-6: $EL_{wait}(LAT, LONG) = 0.3 * EL_{min}(LAT - 1, LONG + 1) + 0.4 * EL_{min}(LAT, LONG + 1) + 0.3 * EL_{min}(LAT + 1, LONG + 1)$

Latitude 7: $EL_{wait}(LAT, LONG) = 0.7 * EL_{min}(LAT, LONG + 1) + 0.3 * EL_{min}(LAT - 1, LONG + 1)$

At each decision point the optimal decision was the one with the lowest cost or expected loss, named EL_{min}. An expected loss difference was calculated for each decision and a mean was calculated for each trial ($EL_{difference}$)[1]. In the equation below, i is the decision point (which is the same number as the longitude) and n is the point at which a final decision was made. A mean was calculated for participant over the 40 trials. Participant's expected loss difference was always a negative number and zero in the case of an optimal decision

$$EL_{difference} = \frac{\sum_{i=1}^{n} EL_{min} - EL_{decision}}{n} \quad (5)$$

The Grid of Expected Losses (Figure 11) shows the expected loss values of all possible participant decisions by storm position. The smallest number in each cell was considered the optimal decision.

[1] 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.

Latitude \ Longitude	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	
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	
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	
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	
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	
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	
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	

Key: wait (W), shelter (S), not take shelter (NS)

Figure 11. Storm location grid with expected loss of decisions.

A one-factor ANOVA was conducted on the expected loss difference with the independent variable forecast format. There was main effect for forecast formats, $F(4,484) = 5.19$, $p<.001$, $\eta^2=.041$. Participants in the percent chance condition had the smallest expected loss difference *(M=-4.10, SD=2.70)*, and those in the watch & warning condition had the largest *(M=-6.65, SD=4.92)*. Planned contrasts showed watch & warning decision quality was significantly worse than in the percent chance $F(1,484) = 19.66$, $p <.0001$ and percent chance range conditions *(M=-5.18, SD=4.11)*, $F(1,484) = 6.84$, $p<.01$. However, participants in the color *(M=-5.75, SD=4.31)* and color + percent chance range *(M=-5.35, SD=3.29)* conditions did not perform significantly better than those in watch & warning. This suggests that decision quality of the status quo, watch & warning, is improved with forecasts that contain numbers.

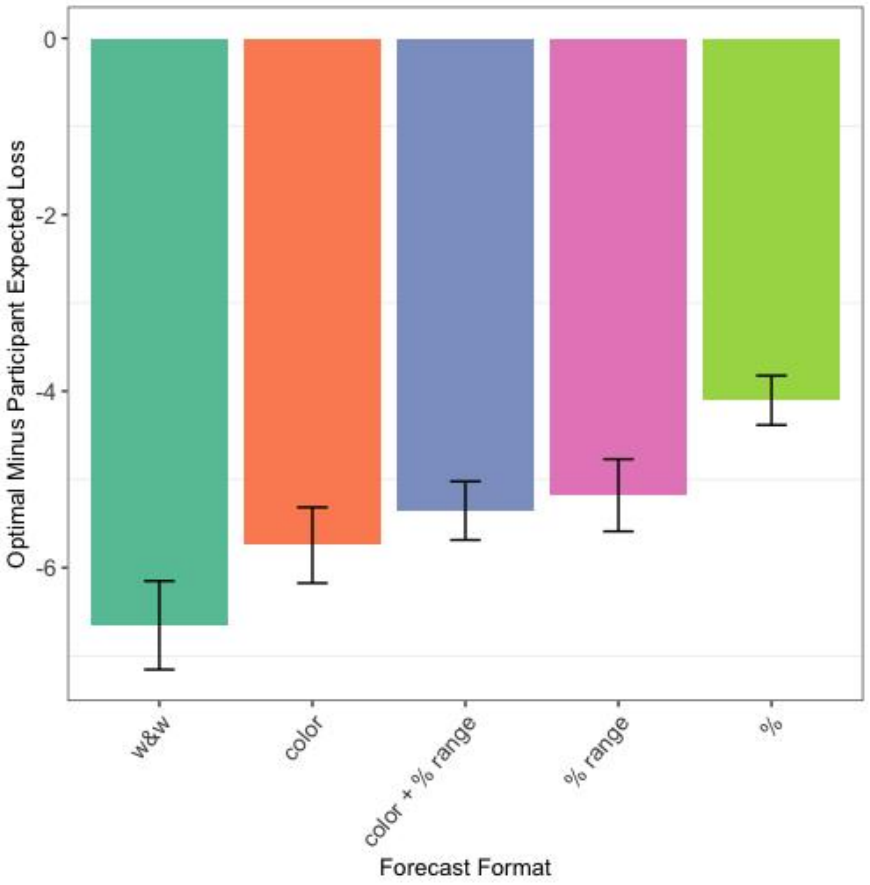

Figure 12. Mean expected loss difference by forecast format. The means closer to zero indicate better decisions because there is little difference in the participant's expected loss values and the optimal expected loss values, on average.

In order to understand whether participants were more cautious (i.e. take shelter more often) with certain forecast formats, the proportion of take shelter decisions was calculated summed across participants' 40 trials. A one-factor ANOVA was conducted on proportion of take shelter decisions with the independent variable forecast format. There was a significant main effect of forecast format, $F(4,484) = 5.19$, $p<.001$, $\eta^2=.041$. Participants in the watch & warning *(M=.37, SD=.21)* had the highest proportion of take shelter decisions and the lowest in the probability condition *(M=.26, SD=.16)*, This was the only contrast that was significant, $F(1,484)$, $p<.0125$. Although not significant, color also lead the highest proportion of take shelter decisions when compared to the other four conditions *(M=0.36, SD=0.18)*.

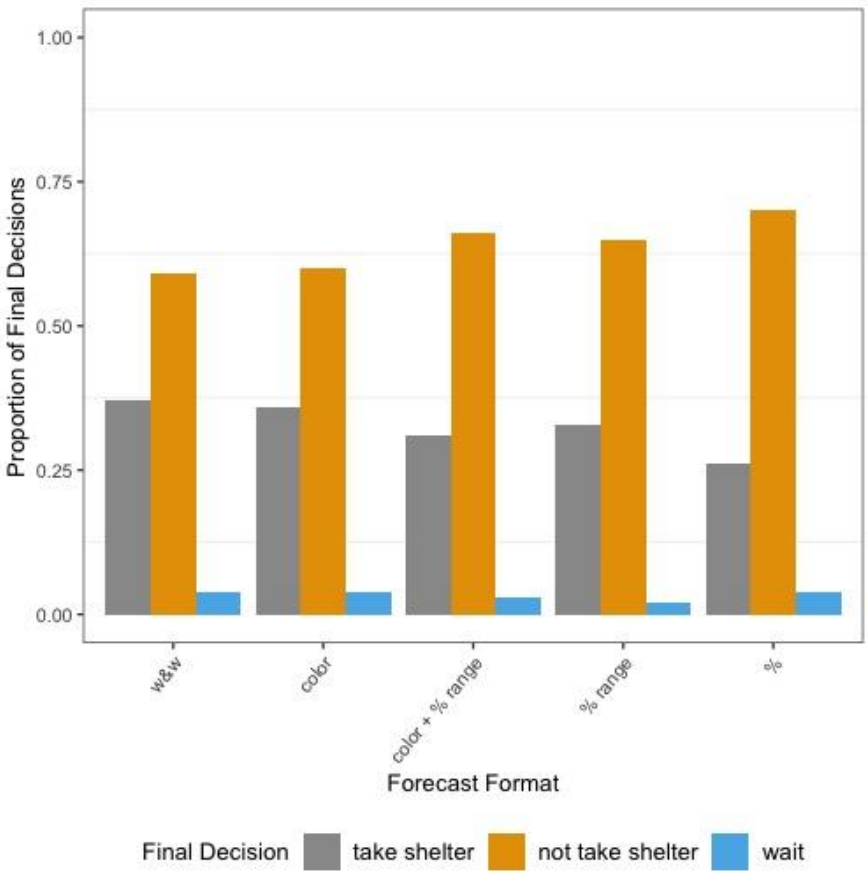

Figure 13. Proportion of final decision out of 40 by forecast format.

The study presented here sought to measure the effect of forecast format on the timeliness of decisions. Timeliness was operationalized as a participant's divergence from the optimal stopping point on a given trial. In order to calculate the timeliness of decisions, the participants stopping point (when they made a final decision) was subtracted from the optimal stopping point (when they should have made a final decision) for each trial. A negative number meant that the participant made a decision prior to the optimal time point; a zero meant they made a decision at the optimal time point; and a positive number meant they made a final decision after the optimal time point. The differences were summed for the participant's 40 trials and divided by 40 to find the mean difference in stopping points.

Participants on average made timely decisions in all forecast formats with the exception of the probability condition, which showed a slight delay beyond optimal stopping. The bar chart of stopping differences (see Figure 14) shows the stopping differences from the optimal decision by forecast format. A one-sample t-test compared the stopping difference in stopping point to zero by forecast format. The stopping in the probability condition *(M = 0.29, SD=1.22)* was significantly greater than zero, $t(92)=2.34$, $p<.05$. However, all other formats were not significantly different from zero, (watch & warning *(M=-0.33, SD=1.81)*; color *(M=-0.05, SD= 1.70)*; color + percent chance range *(M=-0.32, SD=1.67)*; and percent chance range *(M=-0.18, SD=1.50)*) $p>.05$.

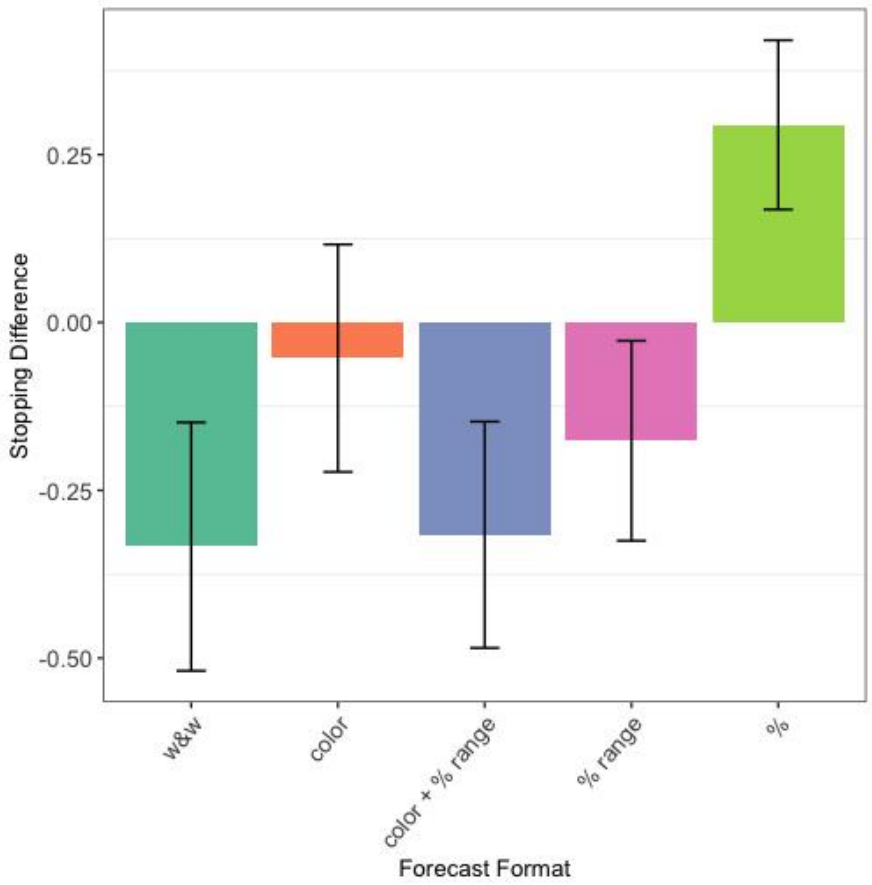

Figure 14. Bar chart mean difference in participant stopping from optimal.

In sum, participants in the probability condition waited longest to make a decision, however they made the best decisions (decision quality) comparatively. Percent chance range showed comparable decision quality to percent chance alone but did not delay significantly beyond optimal stopping. Participants in the watch & warning and color conditions made cautious decisions more frequently than the rest. However, while cautious decisions may be a desirable outcome, decision quality was the poorest in the watch & warning condition.

Discussion

This experiment provides strong evidence that numeric likelihood information is an effective means for communicating forecast uncertainty in a dynamic decision environment in which multiple sequential percent chance forecasts must be evaluated to make a final decision. Participants in the percent chance and percent chance range conditions were most accurate in their interpretation of the forecast both in terms of the likelihood that was being communicated and in terms of less confusion with severity. Participants trusted forecasts with percent chance the most after learning the storm outcome. Furthermore, participants in the percent chance and percent chance range information made the highest quality decisions as shown by their near optimal expected losses.

Participants understanding of likelihood information differed between forecast formats and between ranges of percent chance. All formats showed a tendency to overestimate the likelihood of a tornado hitting their area. Overall, percent chance and percent chance range had likelihood ratings closest to the forecast and showed similar levels of differences from the actual percent chance. This finding suggests that in a real world setting, forecasters can include second-order uncertainty information (e.g. a range such as 0-12%), and thus communicate the ambiguity of the forecast without a detriment to understanding. Ambiguity in forecasting may come from

different forecasts coming from a range of available weather models, each with their own inherent uncertainty, for example. However, including color in the forecast severely degraded understanding of event likelihood even with accompanying percent chances. This was seen in the significantly greater likelihood differences in the color + percent chance range when compared to the percent chance range. In addition, participants' overestimation increased with the actual percent chance. The likelihood differences grew larger with each range (0-12%, 13-24%, 25-40%), resulting in average likelihood differences at higher probabilities of 4.64, 14.47 , 30.64, respectively. Participants estimates showed the greatest overestimation from the actual percent chance in the watch & warning and color-coded forecasts, and it was reflected in the lower quality of their decisions. The magnitude of overestimation was similar for watch & warning and color though likely for different reasons. The watch & warning condition did not provide overt likelihood information but by definition, a tornado warning means that a tornado is imminent and thus, may have implied 100% certainty. When participants saw a warning (or an orange forecast in the case of color) the actual percent chance was between 25% and 40%. When they saw a watch or yellow, it was between 13% and 24%. When they saw no watch or warning, or green, it was between 0 and 12%. Therefore, one might expect especially high estimations in the upper range for watch & warning (i.e. warning) because it may be interpreted as certain and a lower overestimate for color. In fact, the opposite was seen. When orange was shown, the average likelihood ratings for color were higher (*M=61.3*) than when a warning was shown (*M=53.4*) – almost equal to that of color + percent range (*M=52.6*). Another explanation that might account for the pronounced overestimation in both the watch & warning and color conditions is that participants assumed that each color or category represented one third of a scale from impossible to certain, thus markedly increasing likelihood estimates. Under this assumption, likelihood

estimates in the 0-.12, .13-.24, and .25-.40 actual percent chance ranges would have fallen between 0 and 33%, 34% and 66%, and 67% and 100%, respectively. This was true for the 0-12% (watch & warning, *M=30.28*; color, *M=29.03*) and the 13-24% (watch & warning, *M=42.28*; color, *M=40.5*) percent chance ranges. However, this was not the case for 25-40% range (watch & warning, *M=53.58*; color, *M=61.25*) which were lower than would be expected (67-100%). Recall that in the 0-12% and 13-24% ranges, color and watch & warning increased with similar magnitudes. In the upper range however, overestimation in color surpassed that of watch & warning. This suggests that it is not categorization alone that leads to overestimation but rather that it has to do with color in particular. The overestimation in the color condition is in line with previous research findings that showed warnings with color are perceived as more hazardous overall (Braun, Mine, & Silver, 1995). Effects such as these could lead to an increased perception of likelihood (i.e. why likelihood estimates of color-coded forecasts were higher than those with percent chances), and why a condition with the same number of categories as watch & warning would result in different magnitudes of overestimation depending on the color.

In addition, participants had a tendency to misinterpret color-coded likelihood information as saying something about severity in all three operationalizations (severity difference, severity-likelihood correlation, and severity SD). In the absence of explicit likelihood information (as in the watch & warning condition) one might be unable to disentangle the concepts of likelihood and severity and instead assume as one increases so does the other. Color on the other hand was introduced in the instructions as indicating levels of percent chance. Despite the explanation, participants interpreted it as both likelihood and severity. However, with the addition of numeric likelihood (color + percent chance range), severity ratings showed significantly less variability than in the color condition. However, color + percent chance range

showed a lesser tendency for mistaking likelihood for severity, although it was not significantly different from color forecasts. In the case where participants saw color + percent chance range, they most likely adjusted their likelihood ratings down to a number closer to the displayed range, but color still influenced severity. This is compelling evidence that color-coded likelihood information was also interpreted as color-coded severity information. Such a perception is particularly dangerous when the magnitude of the storm is great but the likelihood of a tornado in a particular area is low to medium. Residents may interpret the color to mean even if a tornado did hit, the damage would be minimal or not life threatening and thus fail to take protective action. Percent chance and percent chance range showed the least confusion (largest severity differences, smallest severity-likelihood correlations, and highest severity SDs), suggesting that these formats best allows users to understand that it is likelihood and not severity that is being communicated.

As shown in previous studies, participants made higher quality decisions with numeric likelihood forecasts (percent chance range and percent chance) than with categorical forecasts (i.e. color-coded or watch & warning forecast) (Joslyn & LeClerc, 2013). Initially this may not seem surprising because of the greater precision in the information provided in the percent chance conditions. However, it is important to note that although color and watch & warning included only three categories, the divisions between categories were based on the appropriate decision for the underlying probabilities of the tornado hitting home. When the economically optimal decision was to *not 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, while participants in the percent chance and percent chance range condition

had precise likelihood information, those in the categorical conditions had information tailored to their costs and losses. If participants had chosen to shelter whenever the highest category was forecasted (i.e. warning, orange), their decision quality would have been superior. Instead, participants in the color-coded and watch & warning forecasts chose to take precautionary action more frequently, and often unnecessarily expending resources to avoid a tornado that was unlikely. The perception of high likelihood as shown by participants' overestimation of likelihood in those conditions may have influenced the decision to take shelter more often. It could be that the color yellow or the watch message alone was not sufficiently explicit in its expression of likelihood and led participants to infer the forecasts were more likely, ultimately leading to unwarranted protective action.

While percent chance information proved to be advantageous overall, watch & warning and color-coded forecasts appear to encourage people to take precautionary action. Some may argue this is an advantage that should be exploited. Indeed, color could be a useful communication format when precautionary action should be urged for low probability high magnitude events. Although, there may be a cost in terms of lowered trust that would affect response to subsequent warnings. Indeed, post-outcome trust was significantly lower than the percent chance condition. However, adding percent chance range to the color forecast (color + percent chance range) did show an improvement on color alone with respect to understanding event likelihood and not confusing likelihood with severity. Despite improvements, however understanding never matched that of percent chance and percent chance range, therefore color may be detrimental for use in most forecast communications.

Participants trusted percent chance forecasts significantly more than color-coded forecasts. It may be that trust in color-coded forecasts was diminished because of frequent and

unnecessary decisions to take shelter. Color-coded forecasts may have led participants to believe a tornado was more likely and consequently to make cautious decisions more frequently. Once they learned the outcome and that the cost to take protective action was sometimes unnecessary, their trust may have been lowered.

In some cases of severe weather warning, people delay taking precautionary action and continue to collect information (Comstock & Mallonee, 2005; Schwartz & Howell, 1985). On average, in the experiment people made timely decisions in most formats with the exception of the percent chance condition, but the delay was less than a decision point on average. In the original Schwartz and Howell study (1985) delay (information gathering) was most pronounced in conditions with time pressure as it reduced time to process available information and encouraged participants to seek additional information. One factor in our study that differed from the original Schwartz and Howell study (1985) may have contributed. Our study did not include time pressure. However, as was the case with the 1985 study comparing numeric and graphic displays of storm location, numeric formats encouraged oversampling (waiting for more information) and its effect was much smaller than that of time pressure. It should be noted that percent chance range (unlike percent chance) did not show delay beyond optimal stopping, and decision quality and understanding of event likelihood and severity was comparable to percent chance alone. In future studies the addition of time pressure may reveal sub-optimal stopping that may be affected by forecast format.

This study provides compelling evidence for the benefits to decision making by including numeric probability information in warning forecasts. That said, one might argue that the population under study was college educated and may have had more familiarity with statistical information and therefore had an advantage. However, previous research shows that populations

with low education also benefit from the inclusion of numeric probabilities over deterministic forecasts (Grounds & Joslyn, 2018). However, it is not known if low education groups would benefit in a dynamic decision environment such as the one tested here. Future experiments should investigate individual differences in an environment with continually updated information.

In sum, the results presented here suggest that numeric probability information is an effective format for supporting appropriate response to warning forecasts. In fact, even with multiple forecast updates, including numeric likelihood information in both first and second order uncertainty (e.g. 25%, 25-40%) improved decision making, forecast understanding, and trust. Color-coded and watch & warning forecasts, in an attempt to simplify the communication, may have been detrimental. Participants, left to estimate the likelihood themselves, tended to overestimate the likelihood in both overall and in the highest categories. The perception of higher event likelihood may have negatively affected the decision quality of watch & warning and the trust in color forecasts compared to the percent chance conditions. This suggests that omitting explicit likelihood information leaves an opening for communication error between scientists/public officials and the general public.

References

Ash, K. D., Schumann, R. I., & Bowser, G. C. (2014). Tornado warning trade-offs: Evaluating choices for visually communicating risk. *Weather Climate and Society, 6*(1), 104-118. doi:10.1175/WCAS-D-13-00021.1

Borade, A. B., Bansod, S. V., & Gandhewar, V. R. (2008). Hazard perception based on safety words and colors: An indian perspective. *International Journal of Occupational Safety and Ergonomics, 14*(4), 407-416. doi:10.1080/10803548.2008.11076777

Braun, C. C., Kline, P. B., & Silver, N. C. (1995). The influence of color on warning label perceptions. *International Journal of Industrial Ergonomics,* (15), 179-187.

Braun, C. C., Kline, P. B., & Clayton Silver, N. (1995). The influence of color on warning label perceptions. *International Journal of Industrial Ergonomics, 15*(3), 179-187. doi:10.1016/0169-8141(94)00036-3

Chapanis, A. (1994). Hazards associated with three signal words and four colours on warning signs. *Ergonomics, 37*(2), 265-275. doi:10.1080/00140139408963644

Comstock, R. D., & Mallonee, S. (2005). Comparing reactions to two severe tornadoes in one oklahoma community. *Disasters, 29*(3), 277-287. doi:10.1111/j.0361-3666.2005.00291.x

Enten, H. (2015, -01-27T20:41:26+00:00). How meteorologists botched the blizzard of 2015. Retrieved from https://fivethirtyeight.com/features/how-meteorologists-botched-the-blizzard-of-2015/

Gigerenzer, G., Hertwig, R., Van, D. B., Fasolo, B., & Katsikopoulos, K. V. (2005). “A 30% chance of rain tomorrow”: How does the public understand probabilistic weather forecasts? *Risk Analysis, 25*(3), 623-629. doi:10.1111/j.1539-6924.2005.00608.x

Goldman, M., & Rao, J. M. (2017). Optimal stopping in the NBA: Sequential search and the shot clock. *Journal of Economic Behavior & Organization, 136*, 107-124. doi:10.1016/j.jebo.2017.02.012

Grounds, M. A., & Joslyn, S. L. (2018). Communicating weather forecast uncertainty: Do individual differences matter? *Journal of Experimental Psychology: Applied, 24*(1), 18-33. doi:10.1037/xap0000165

Hellier, E., Tucker, M., Kenny, N., Rowntree, A., & Edworthy, J. (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. *Journal of Patient Safety, 6*(3), 158. doi:10.1097/PTS.0b013e3181eee157

Hershman, R. L., & Levine, J. R. (1970). Deviations from optimum information-purchase strategies in human decision-making. *Organizational Behavior and Human Performance, 5*(4), 313-329. doi:10.1016/0030-5073(70)90023-1

Joslyn, S. L., Nadav□Greenberg, L., Taing, M. U., & Nichols, R. M. (2009). The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Applied Cognitive Psychology, 23*(1), 55-72. doi:10.1002/acp.1449

Joslyn, S., & Leclerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science, 22*(4), 308-315. doi:10.1177/0963721413481473

Joslyn, S., Pak, K., Jones, D., Pyles, J., & Hunt, E. (2007). The effect of probabilistic information on threshold forecasts. *Weather and Forecasting, 22*(4), 804-812. doi:10.1175/WAF1020.1

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications, 17*(2), 180-195. doi:10.1002/met.190

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment, 49*, 81.

Kline, P. B., Braun, C. C., Peterson, N., & Silver, C. (1993). The impact of color on warnings research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 37*(14), 940-944.

LeClerc, J., & Joslyn, S. (2015). The cry wolf effect and Weather□Related decision making. *Risk Analysis, 35*(3), 385-395. doi:10.1111/risa.12336

Lesch, M. F., Rau, P. P., Zhao, Z., & Liu, C. (2009). A cross-cultural comparison of perceived hazard in response to warning components and configurations: US vs. china. *Applied Ergonomics, 40*(5), 953-961. doi:10.1016/j.apergo.2009.02.004

Mayhorn, C. B., Wogalter, M. S., Bell, J. L., & Shaver, E. F. (2004). What does code red mean? *Ergonomics in Design: The Quarterly of Human Factors Applications, 12*(4), 12-14. doi:10.1177/106480460401200404

McPherson, R. A., Brooks, H., Greene, S., Purcell, D., Tarhule, A., Thomas, R., & Klockow, K. (2013). In McPherson R. A., Brooks H., Greene S., Purcell D., Tarhule A. and Thomas R.(Eds.), *Spatializing tornado warning lead-time: Risk perception and response in a spatio-temporal framework* ProQuest Dissertations Publishing.

Mileti, D., & Fitzpatrick, C. (1991). Communication of public risk: Its theory and its application. *Sociological Practice Review, 2*(1), 20-28.

Morss, R., Demuth, J., & Lazo, J. (2008). Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Weather and Forecasting, 23*(5), 974-991. doi:10.1175/2008WAF2007088.1

Nagele, D., & Trainor, J. (2012). Geographic specificity, tornadoes, and protective action. *Weather Climate and Society; Weather Clim.Soc., 4*(2), 145-155. doi:10.1175/WCAS-D-11-00047.1

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

Paul, B., & Stimers, M. (2012). Exploring probable reasons for record fatalities: The case of 2011 joplin, missouri, tornado. *Natural Hazards; Journal of the International Society for the Prevention and Mitigation of Natural Hazards, 64*(2), 1511-1526. doi:10.1007/s11069-012-0313-3

Radford, L., Senkbeil, J. C., & Rockman, M. (2013). Suggestions for alternative tropical cyclone warning graphics in the USA. *Disaster Prevention and Management: An International Journal, 22*(3), 192-209. doi:10.1108/DPM-06-2012-0064

Rashid, R., & Wogalter, M. S. (1997). Effects of warning border color, width, and design on perceived effectiveness. *Advances in Occupational Ergonomics and Safety,* , 455-458.

Savelli, S., & Joslyn, S. (2013). The advantages of predictive interval forecasts for non□expert users and the impact of visualizations. *Applied Cognitive Psychology, 27*(4), 527-541. doi:10.1002/acp.2932

Schwartz, D. R., & Howell, W. C. (1985). Optional stopping performance under graphic and numeric CRT formatting. *Human Factors, 27*(4), 433-444. doi:10.1177/001872088502700407

Sherman-Morris, K., Antonelli, K., & Williams, C. (2015). Measuring the effectiveness of the graphical communication of hurricane storm surge threat. *Weather Climate and Society, 7*(1), 69-82. doi:10.1175/WCAS-D-13-00073.1

Slovic, P., & Lichtenstein, S. (1971). Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*(6), 649-744. doi:10.1016/0030-5073(71)90033-X

Tak, S., & Toet, A. (2014). *Color and uncertainty: It is not always black and white* The Eurographics Association. doi:10.2312/eurovisshort.20141157

US Department of Commerce, NOAA.Watch/warning/advisory definitions. Retrieved from https://www.weather.gov/lwx/WarningsDefined

Wogalter, M. S., Magurno, A. B., Carter, A. W., Swindell, J. A., Vigilante, W. J., & Daurity, J. G. (1995). Hazard associations of warning header components. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 39*(15), 979-983. doi:10.1177/154193129503901503

Wogalter, M. S., Conzola, V. C., & Smith-Jackson, T. (2002). Research-based guidelines for warning design and evaluation. *Applied Ergonomics, 33*(3), 219-230. doi:10.1016/S0003-6870(02)00009-1

Appendix A

General Instructions

Welcome. Please read over the consent form at your desk. Sign the very bottom line if you agree. Please do not begin the game. We will read over the instructions together.

Has anyone not given me the consent form?

(once all forms are collected) Take a moment to silence your phones if you have not done so already.

Go ahead and enter the atmospheric sciences courses you have taken, your age, sex, and hit submit. Wait to proceed further in the game.

You should now be on the “background screen.” I will read over the instructions. Please follow along as I read but do not click ahead until I say “next.”

Background

Tornadoes are dangerous weather phenomena involving high-speed circulating winds as part of strong storms. The circulating winds typically create vortexes, or moving columns of air, that stretch from the ground to the clouds. Although tornadoes have been observed throughout the world, they are most common in the central United States in an area called "Tornado Alley".

Tornadoes vary in size and wind speed, and are responsible for significant damage and loss of many lives in "Tornado Alley" every year. Tornadoes can form quickly and move sporadically, making prediction and warnings difficult. The National Weather Service monitor real-time weather data to provide forecasts to residents. Although forecasts for tornadoes have improved significantly, they remain difficult to accurately predict.

The most violent tornadoes (260+ mph) can level and blow away almost any house and its occupants. However, extremely violent tornadoes are very rare. Most tornadoes are weaker and can be survived by taking protective action. You can move to a tornado shelter, the basement of a house or, on the lowest level of a house, small center room (like a bathroom or closet), under a stairwell, or in an interior hallway with no windows. If no action is taken, you risk being injured by dangerous flying debris as a result of even weak tornadoes (73-112 mph). NEXT

Training

You are about to participate in a simulation in which you will be asked to make decisions about what to do as a series of storms approach a house where you are located. These storms have the potential to produce tornadoes. For each storm, you will decide to do one of three things:

1. **Wait** for more information

2. **Take Shelter** in a tornado shelter near your house
3. **Not Take Shelter**

The difference between "wait" and "not take shelter" is that "not shelter" is a final decision.

Such decisions are important in real life because a tornado strike can cause significant injury and possible loss of life. In the simulation, the consequences of your decisions are reflected in a point system that you can turn-in at the end of the experiment for a possible cash reward.

NEXT

Training

You will experience 40 storms, or rounds. Each round begins with a new storm system moving west to east, from the same distance, toward the home in which you are located. For the purposes of this simulation, you can assume all tornadoes produced by these storm systems have wind speeds of 90 to 112 miles per hour.

For each round, you will make seven decisions. At each of the seven decision points you will receive new information about the storm as it progresses toward the home. After you examine the information, you will decide whether to:

- Wait
- Not take shelter (final)
- Take shelter (final)

When you choose to wait at decision points 1-6, the storm moves forward, you will get additional information, and you will be allowed to make another decision. However, decisions to take shelter or not take shelter are final. Making a final decision ends that round (although you will get to see the rest of the information, answer a few more questions, and see whether or not a tornado hits home).

NEXT

(Note: The text between the following red lines is specific to each experimental condition. Information about the watch & warning condition is included as an example.)

Training - Information

At each decision point you will receive a forecast like the one below, telling you whether the home is in an area under tornado watch or warning, or neither.

Your area is currently not under a watch or warning.

NEXT

Or… Your area is currently under a tornado watch.

Watch means that tornadoes are possible in or near the watch area.

NEXT

Or… Your area is currently under a tornado warning.

Warning means that a tornado has been sighted or indicated nearby and may enter the warning area.

NEXT

Training - The Costs & Penalty

You will have a budget of 24,000 to start. Your goal is to finish the 40 rounds with the highest possible point budget.

Costs:

- The cost to **take shelter begins at 303 points**, because taking shelter can be costly in terms of effort, time, and sometimes money. The cost of taking shelter increases as the storm approaches home, to reflect the increasing danger of being caught in a vulnerable position when a tornado strikes. If you take shelter and a tornado hits the house, you will be safe.
- **Waiting** has **no cost** for decision points 1-3.
  For decision points 4-7 there is a **20-point** cost for every "wait" decision.
- **Not Take Shelter** has **no cost**

Penalties:

- If a tornado hits home and you have not chosen to take to shelter, you will be penalized **1500** points. However, if a tornado does not strike the house, you will not be penalized, regardless of your choice.

First, you will have 10 practice rounds to fully learn the game. BEFORE YOU BEGIN a few administrative comments…

When you are done, please raise your hand. I will provide you with a debriefing sheet and an online survey to complete. I will record your remaining points. You will receive $1 if you have at least 13,380 point remaining. You will receive a dollar for every additional 1500 points remaining. You will get 1.0 course credit regardless of your point budget. Please let me know if you have any questions. Go ahead and click next to begin.

Condition-Specific Instructions

(Note: The text below should be read as part of the general instructions. Include in the general instructions above between the two red lines identical to those shown below.)

Color Condition

Training - Information

At each decision point you will receive a forecast, like the one below, with a color.

(Next screen)

(Next screen)

Warmer colors indicate higher likelihood.

NEXT

Percent Chance Ranges Condition

Training - Information

At each decision point you will receive a forecast, like the one below, telling you the likelihood range of a tornado hitting home.

Chance of tornado hitting your area:

0-12%

(Next screen)

Chance of tornado hitting your area:

13-24%

(Next screen)

Chance of tornado hitting your area:

25-40%

The percentage is the chance from 0 to 100% that the tornado will hit home by the end of the round (seven decisions).

NEXT

Color + Percent Chance Ranges Condition

Training - Information

At each decision point you will receive a forecast, like the one below, with a color value and the percent chance range of the tornado hitting your area.

Chance of tornado hitting your area:

0% to 12%

(Next screen)

Chance of tornado hitting your area:

13% to 24%

(Next screen)

Chance of tornado hitting your area:

25% to 40%
Warmer colors indicate higher likelihood.

The percentage is the chance from 0 to 100% that the tornado will hit home by the end of the round (seven decisions).

(Next screen)

Percent Chance Condition

Training - Information

At each decision point you will receive a forecast, like the one below, telling you the likelihood of a tornado hitting home.

Chance of tornado hitting your area: 6%.

(Next screen)

Chance of tornado hitting your area: 19%.

(Next screen)

Or…Chance of tornado hitting your area: 33%.

The percentage is the chance from 0 to 100% that the tornado will hit home by the end of the round (seven decisions).

(Next screen)

Watch & Warning Condition

Training - Information

At each decision point you will receive a forecast like the one below, telling you whether the home is in an area under tornado watch or warning, or neither.

Your area is currently **not** under a watch or warning.

(Next screen)

Your area is currently under a tornado watch.

Watch means that tornadoes are possible in or near the watch area.

(Next screen)

Your area is currently under a tornado warning.

Warning means that a tornado has been sighted or indicated nearby and may enter the warning area.

(Next screen)

Appendix B

Participant Debriefing

Thank you for participating in our experiment! It has long been recognized that display format can have a profound, though task-dependent, influence on human performance. Display format refers to the way that the likelihood of a certain event is displayed. Because many decisions in the real world are made under time pressure such as, for instance, deciding to stay, wait, or evacuate after being informed of the risk of an incoming tornado, we were interested in comparing people's performance in a game given different types of display formats under different time constraints. If you have any questions about your participation in our experiment, do not hesitate to contact our research office at gala@uw.edu.

Appendix C

Participant Debriefing Survey

How did you make your decisions to wait, shelter, or not take shelter?

What information did you use to make your decisions to wait, shelter, or not take shelter?

Did you develop a strategy or a rule in playing the game? What was it?

You received information about each storm. In your own words, what did the information tell you?

You estimated the severity of the tornado damage. What information did you use to answer the question?