Colorful Language: Investigating Public Interpretation of the Storm Prediction Center Convective Outlook

SEAN ERNST,[a] JOE RIPBERGER,[a] MAKENZIE J. KROCAK,[a] HANK JENKINS-SMITH,[a] AND CAROL SILVA[a]

[a] *Center for Risk and Crisis Management, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 5 January 2021, in final form 8 July 2021)

ABSTRACT: Although severe weather forecast products, such as the Storm Prediction Center (SPC) convective outlook, are much more accurate than climatology at day-to-week time scales, tornadoes and severe thunderstorms claim dozens of lives and cause billions of dollars in damage every year. While the accuracy of this outlook has been well documented, less work has been done to explore the comprehension of the product by nonexpert users like the general public. This study seeks to fill this key knowledge gap by collecting data from a representative survey of U.S. adults in the lower 48 states about their use and interpretation of the SPC convective outlook. Participants in this study were asked to rank the words and colors used in the outlook from least to greatest risk, and their answers were compared through visualizations and statistical tests across multiple demographics. Results show that the U.S. public ranks the outlook colors similarly to their ordering in the outlook but switches the positions of several of the outlook words as compared to the operational product. Logistic regression models also reveal that more numerate individuals more correctly rank the SPC outlook words and colors. These findings suggest that the words used in the convective outlook may confuse nonexpert users, and that future work should continue to use input from public surveys to test potential improvements in the choice of outlook words. Using more easily understood words may help to increase the outlook's decision support value and potentially reduce the harm caused by severe weather events.

SIGNIFICANCE STATEMENT: The Storm Prediction Center's Convective Outlook, though originally designed for expert users, has become popular as a communication tool for the general public through broadcast and social media. We wanted to identify a baseline for how well people understand the words and colors used in the outlook to communicate risk, using a series of survey questions issued to U.S. adults. Our results suggest that the outlook words do not clearly communicate risk to public users, and that future iterations of the outlook should use terms that are easily sequentially ordered. Future work should seek to identify terms that are more easily understood by experts and nonexperts alike.

KEYWORDS: Social Science; Convective storms; Forecasting techniques; Operational forecasting; Communications/ decision making; Decision support

1. Introduction

The Storm Prediction Center (SPC) convective outlook is one of the oldest continuous severe weather forecasts, having existed in one form or another since 1955 (Corfidi 1999; Hitchens and Brooks 2012). The outlook forecasts the likelihood of severe weather (including tornadoes, convective wind, and hail) within 25 miles of any given point across the United States for the 1–8-day period, presenting those probabilities in numerical form as well as translating them into a five-tier scale of words and matched colors (for the day 1–3 period). The outlook product was originally designed for internal government use, but in the last decade has become widely referenced across social media and television (Cappucci 2020). Some of this increased visibility may originate from the increasing value individuals place on advance warning of severe weather, as the cost of severe weather disasters has soared from a 5-yr average of \$1.3 billion (U.S. dollars, adjusted for inflation) in 1984 to over \$15 billion in 2020 (NCDC 2020). Forecasts like the convective outlook do not inherently have value; however, as Murphy (1993) describes that value is generated through the decisions forecast users make using the information contained within the forecast product. Though many studies have investigated the forecast quality and accuracy of the outlook (Hitchens and Brooks 2012, 2014, 2017; Hitchens et al. 2013), and some studies have investigated the value that emergency managers generate from the outlook (e.g., Ernst et al. 2018), there has been a lack of research into whether the general public is able to glean information from the outlook from which they can generate value. This study seeks to take the first steps toward bridging this knowledge gap by investigating how accurately members of the public rank the risk words and colors used to communicate risk in the outlook. The results of this study may further indicate ways to improve the design of the outlook to better communicate threats to users, potentially allowing them to prepare themselves for severe weather impacts well in advance of their arrival.

a. History of the SPC convective outlook

The first regularly issued version of the SPC convective outlook was released in 1973, highlighting areas of Moderate and High risk for severe weather, as well as a lower tier of Slight risk after 1974 (Corfidi 1999; Hitchens and Brooks 2012). The outlook was originally only issued for the day 1 period, but outlooks for the day 2, 3, and 4–8 periods were introduced in 1986, 2000, and 2005, respectively (Edwards and Ostby 2015).

Corresponding author: Sean Ernst, Sean.Ernst@ou.edu

DOI: 10.1175/WAF-D-21-0001.1

These outlooks were initially issued based on subjective forecaster interpretation of risk but were married to probabilistic forecasts of storm coverage beginning in 2003 (Fig. 1).

Yet another major change to the outlook was made in 2014, when two new categories were added to the outlook to increase its detail at lower risk levels (Edwards and Ostby 2015). This decision was driven in part by concerns from emergency managers, who wanted the multiple probability thresholds in the probabilistic outlook encompassed by the Slight risk category (see Fig. 1) to be shown in the categorical outlook as well (P. Marsh 2020, personal communication). Emergency managers and the head of FEMA at the time, Craig Fugate, were also adamant that any changes to the outlook should not alter the Moderate or High categories due to their use as thresholds for emergency management operations. After administering a public survey on their website regarding the new categories, the SPC made the decision to replace ''See Text'' with ''Marginal'' and the higher probabilities of ''Slight'' with ''Enhanced Slight.'' Enhanced Slight was then shortened to ''Enhanced'' in the final design, which was applied to the day 1, 2, and 3 outlooks in 2014 (see Fig. 2; P. Marsh 2020, personal communication).

Though the SPC moved forward operationally with the new categorical outlook design, SPC leadership anticipated that controversy would emerge over the new names in the outlook (P. Marsh 2020, personal communication). In anticipation of concerns over the wording, the SPC worked to wed colors and numerical order to the categorical levels, and developed a graphic describing the expected impacts that each category suggested (see Fig. 3). This color and word scale has since become the de facto design used by multiple government organizations, for example the excessive rainfall outlook issued by the Weather Prediction Center (WPC 2020). Though no changes have been made to the design of the operational product since the new scale was adopted, the SPC is currently working on updating their products as part of the Forecasting A Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) initiative. New forecast products are currently undergoing accuracy testing at the Hazardous Weather Testbed in Norman, Oklahoma, but the reception of the outlook by members of the public is still uncertain and needs additional research beyond postevent service assessments (Pietrycha and Fox 2004; NOAA 2011; Stough et al. 2012; Ernst et al. 2018).

b. Public interpretation of the SPC convective outlook

In recent years the SPC has sought to understand how its core partners interpret the convective outlook (P. Marsh 2020, personal communication), and a suite of papers have investigated the verification of the accuracy of the outlook (Hitchens and Brooks 2012, 2014, 2017; Hitchens et al. 2013; Herman et al. 2018), but these efforts have not yet sought to define how the outlook is interpreted by the general public. More recent studies have started to shift the focus toward public interpretation, though their attention focuses on comparing different presentations of the outlook. For example, Williams et al. (2020) studied whether consistency across SPC, National Weather Service (NWS), and television broadcaster presentations of the convective outlook impact the public's understanding of the graphic. Other studies have investigated user interpretation of similar outlook graphics, such as the Climate Prediction Center climate outlook, finding that the colors and legends currently used in the visual product can generate confusion (Gerst et al. 2020). This confusion can be made even worse for non-English speaking users, as identified by Trujillo-Falcón et al. (2021). Overall, however, there is a gap in research into how members of the public interpret the words and colors used to communicate risk in the SPC outlook.

Day 1 Probability to Categorical Outlook Conversion

(SIGNIFICANT SEVERE area needed where denoted by hatching - otherwise default to next lower category)

Outlook Probability	TORN	WIND	HAIL
2%	SEE TEXT	NOT USED	NOT USED
5%	SLGT	SEE TEXT	SEE TEXT
10%	SLGT	NOT USED	NOT USED
15%	MDT	SLGT	SLGT
30%	HIGH	SLGT	SLGT
45%	HIGH	MDT	MDT
60%	HIGH	HIGH	MDT

FIG. 1. The probabilistic breakdown of the original Day 1 Convective Outlook for tornado, wind, and hail threats.

Though public SPC outlook interpretation is less well understood, more is known about how members of the public interpret tornado warning information. Studies of the level of public tornado warning comprehension in the United States vary in their findings, ranging from participant comprehension as high as the 90% range (Schultz et al. 2010) to as low as 47% of participants in some regions and racial groups (Powell and O'Hair 2008, Mason and Senkbeil 2015). Additionally, multiple studies have identified interactions between warning comprehension and demographics, with region, race, age, gender, ethnicity, and education all being found to have varying effects on warning comprehension (Powell and O'Hair 2008; Jauernic and Van Den Broeke 2017; Allan et al. 2019; Ripberger et al. 2019). Overall, white, highly educated, middle-aged individuals living in the tornado-prone Great Plains region were found to have the highest tornado warning comprehension. However, as Ripberger et al. (2020) show, notable differences exist between predicted and observed tornado warning comprehension when modeled by NWS county warning area, which may be due to untested demographic disparities. These findings suggest that the links between demographic groups and interpretation of the SPC outlook should also be investigated, as they may reveal communication failures with known vulnerable populations that should be addressed. Further, these findings suggest that we should seek to identify any links between how weather aware or able to understand the tornado warning and watch system members of the public are and their ability to correctly interpret the SPC convective outlook, as this could suggest patterns of weather information comprehension among members of the public.

Another variable that may impact individual SPC outlook comprehension is numeracy, defined by Cokely et al. (2012) as the ability to use mathematical skills to understand and reason with probabilities. Numeracy is strongly correlated with education level, but even individuals in highly educated professional

Day 1 Outlook Probability	TORN	WIND	HAIL
2%	MRGL	Not Used	Not Used
5%	SLGT	MRGL	MRGL
10%	ENH	Not Used	Not Used
10% with Significant Severe	ENH	Not Used	Not Used
15%	ENH	SLGT	SLGT
15% with Significant Severe	MDT	SLGT	SLGT
30%	MDT	ENH	ENH
30% with Significant Severe	HIGH	ENH	ENH
45%	HIGH	ENH	ENH
45% with Significant Severe	HIGH	MDT	MDT
60%	HIGH	MDT	MDT
60% with Significant Severe	HIGH	HIGH	MDT

FIG. 2. The probabilistic breakdown of the current Day 1 and 2 Convective Outlooks for tornado, wind, and hail threats. Note the inclusion of the ENH and MRGL, or Enhanced and Marginal, tiers.

groups were found to vary in their levels of numeracy (Cokely et al. 2018). Risk judgements and general decision-making efficacy are also related to numeracy, likely due to the importance of statistical literacy to the kind of inductive reasoning tested in decision-making experiments (Cokely et al. 2012; Allan 2018). As interpreting the SPC convective outlook product involves linking words and colors to levels of personal risk, more numerate individuals may be better able to interpret the risk conveyed in the outlook.

Further studies have investigated the use of color in forecast products, though their findings do not agree on any one risk communication solution. Lipkus and Hollands (1999) suggested that risk information presented as numbers alone is more difficult for individuals to process, while presenting that information in graphical form can hold a viewer's attention better and improve their ability to process information. However, individual interpretations can vary greatly across products and presentations. While rainbow-colored storm surge forecasts have been found to be preferred over monochromatic graphics by members of the public, rainbow-colored radar precipitation graphics are correctly interpreted less often than monochromatic graphics (Bryant et al. 2014; Morrow et al. 2015). Respondents in Bryant et al. (2014) suggested that the rainbow color scale used for radar reflectivity had too many colors, but that they wanted more colors in the monochromatic scales to enunciate areas of higher precipitation. Combined, these studies suggest that the spectral scale used to convey risk in the outlook may be confusing for users, though the lower number of colors used in the SPC product may counteract the issues found in previous studies for graphics of this design.

c. *Study objectives*

As the public exposure of the SPC convective outlook continues to increase, these questions about how nonexpert users interpret the outlook, and whether some groups of people are less able to interpret the outlook product, will need answers if the outlook is to be of value to the public. To this end, this exploratory study seeks to answer two primary questions; first, do members of the U.S. public correctly order from lowest to

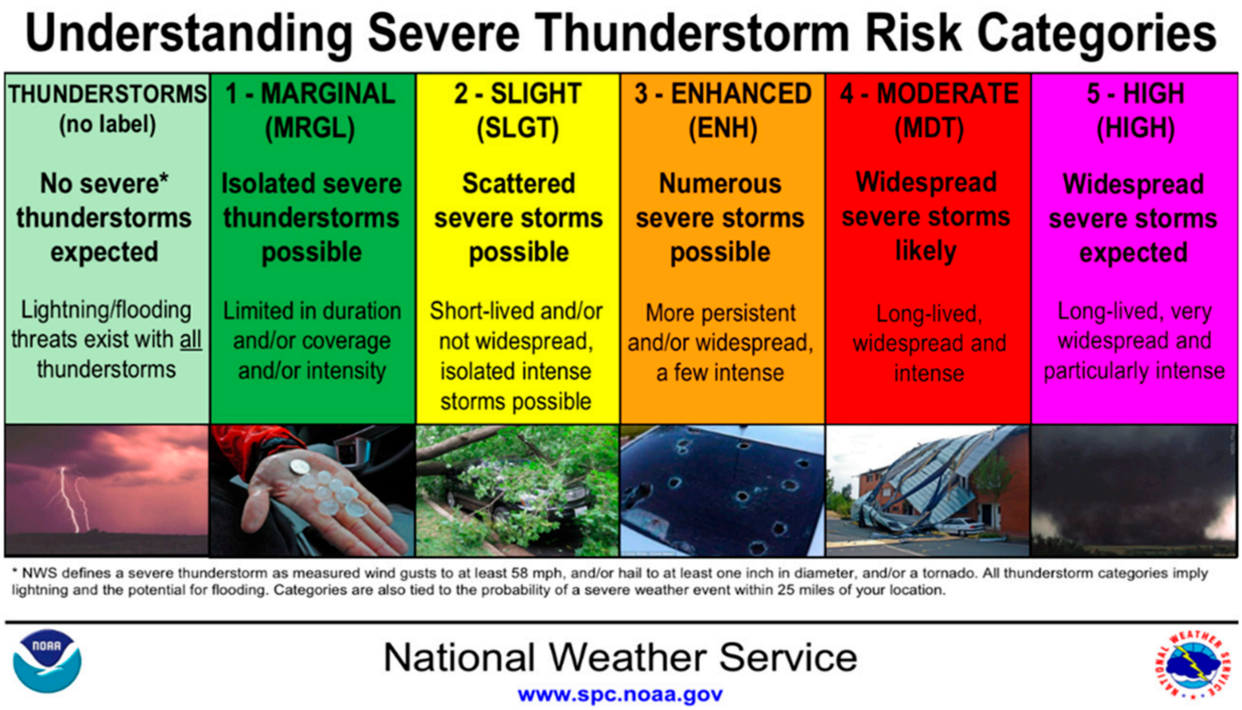

FIG. 3. The SPC's online guide to the meaning of each tier of the convective outlook. Note that the image closely links the words and colors of each outlook tier, also adding numbers and coverage word descriptions to each level.

highest risk level the words and colors associated with the SPC convective outlook? And second, do some commonly defined demographic groups differ in their ability to interpret the convective outlook as compared to other groups? In the following sections, we will discuss the design of the survey we used to collect data for the study, the statistical techniques we used to analyze these data, and finally present our answers to the two exploratory questions we have posed.

2. Data and measures

We used the 2019 Severe Weather and Society Survey (WX19), an annual survey of contiguous United States adults over age 18, to collect data for this study. This survey was developed at the University of Oklahoma by the Center for Risk and Crisis Management (CRCM) and administered to a demographically representative sample that was provided by Qualtrics. Participants were contacted through e-mail and dynamically sampled to generate a representative sample based on U.S. Census data (Table 1, Silva et al. 2019). WX19 sampled 3006 adults, of which 51.3% were male and 48.7% were female. Data and results summaries from WX19, as well as the previous editions of the WX survey, can be accessed online at (https://dataverse.harvard.edu/dataverse/wxsurvey).

For this study, we included a pair of single-issue questions in the WX19 survey that asked participants to order the words and the colors used in the SPC convective outlook from least to greatest risk. Participants were not shown any images of the SPC outlook and were only asked to rank the words and colors from least to greatest risk based on their own perceptions (see Table 2). These questions sought to assess how well individuals interpret the SPC outlook, as well as identify which words or colors were the most problematic in individual's interpretations. It is important to note that the default ordering of the words and colors presented to participants for these questions was randomized, such that no single default order would systematically bias the dataset if participants skipped over the ordering questions. All participants were also prompted with both ordering tasks, with the word ordering task followed by the color task.

Data were also collected from a series of questions that helped define participant's demographic information (Table 2). First, numeracy was estimated using the Berlin Numeracy Test (BNT; Cokely et al. 2012). This test first uses four multiple-choice questions to measure a respondent's statistical numeracy skills. The version used for this study, the BNT-S, includes three additional items adapted from Schwartz et al. (1997) that increase the sensitivity of the measure for less skilled or educated individuals (see the appendix). The BNT-S is an adaptive test, where participants are presented different questions based on their ability to correctly answer each prior question. This adaptive design results in a much faster, less intensive test of numeracy that has been found to be comparably accurate to participant answers when completing the full BNT (Cokely et al. 2012). The BNT and BNT-S items have been shown to exhibit convergent validity with other measures of numeracy and cognitive ability, with statistically significant correlations at the $p <$ 0.01 level with the Lipkus et al. (2001) 11 item numeracy test,

TABLE 1. The demographic breakdown of survey participants as compared to the most recent U.S. Census.

	U.S. adult population (%)	Participants (%)
Gender		
Female	51.3	51.3
Male	48.7	48.7
Age		
18–24	12.0	12.0
25–34	18.0	18.2
35–44	16.3	16.3
45–54	16.4	16.3
55–64	16.7	16.7
65 and up	20.6	20.5
Ethnicity		
Hispanic	16.3	16.4
Non-Hispanic	83.7	83.6
Race		
White	77.9	77.9
Black or African American	13.0	12.8
Asian	5.9	5.9
Other race	3.2	3.4
NWS region		
Eastern	31.6	32.0
Southern	27.1	26.5
Central	20.7	20.9
Western	20.6	20.6

Raven's Advanced Progressive Matrices (RAPM), and the Cognitive Reflection Test (CRT, see Table 3 in Cokely et al. 2012). Further, the test has been shown to provide predictive validity for individual risk understanding, with the test improving the amount of variance explained by statistical models explaining risk understanding that include the RAPM and CRT tests. This version of the BNT has been used extensively in studies of health risk but has only been used in a few previous studies of weather and climate decision-making and risk interpretation, including Allan et al. (2017) and Cho et al. (2021), suggesting more research with numeracy in this field is needed. For this study, the Cronbach's alpha for the BNT-S test was measured to be 0.67, which is comparable to previous studies (Cokely et al. 2012) but modest in comparison to alpha values that we often see for other scales in the social and behavioral sciences. It indicates that there is not a high degree of consistency from item to item on the scale, which suggests that there is more to the numeracy construct than is being assessed by the relatively simple scale we use in this study. Nevertheless, this measure is considered to be one of the best measures for numeracy at this time. Further information and research exploring the use and reliability of the BNT can be found online at RiskLiteracy.org.

We also included a measure of how closely participants follow updates on the weather, which we used as a proxy for how engaged a participant is with weather information. We operationalized this concept by asking participants on a 5-point Likert scale to indicate whether they followed the weather closely. We believe that closely following the weather is a reasonable proxy for weather engagement, but we acknowledge that engagement, like numeracy, is a multidimensional concept that we cannot fully capture with a single survey question.

Finally, participants' objective understanding of the difference between tornado watches and warnings was collected in the survey. Participants were tested with one of two questions, which contained a prompt describing either a tornado warning or a tornado watch. Participants who correctly chose which product the prompt described were then given a dummy variable value of one.

We also captured the amount of time participants spent on the page displaying the two SPC rank order questions. The time in seconds spent on the page is used as a proxy for whether or not participants took the time necessary to meaningfully respond to the ranking questions. Checks like these can help to identify and account for measurement error that arises when participants speed through the survey without carefully reading and responding to all the questions. We also tested a reading comprehension question where we asked participants to ignore the text at the beginning of the question and click a blue dot on the page and measured the total time participants spent answering the complete survey, but both measures had similar results to the time spent on the page measure and thus were not included here. The range of times recorded was extremely large, so recorded times were broken down by quartile group for analysis in our model.

3. Methods and data analysis

a. Public interpretation of the convective outlook

Using these measures, the first step of the analysis was to characterize the data by displaying, in order, how many respondents chose each of the five risk words for the five possible risk levels. Next, the sizes of each ordering group (e.g., how many respondents ranked the risk words as "Marginal, Slight, Enhanced, Moderate, High") were compared. The results of these grouping efforts were developed into a series of visualizations, to aid in interpretation. This process was then repeated for the color ranking data. These visualizations were intended to identify how participants ranked the outlook words and colors, to identify how easily interpreted both risk communication devices in the SPC outlook are by the public.

b. Demographics versus convective outlook interpretation

To identify whether SPC outlook interpretation ability varied across demographic groups, we built a pair of logistic models to compare whether participants ranked the outlook words and colors in the way they are ranked in the outlook product with the independent variables of interest to this study. To do this, participants were divided into two groups for each ranking question, based on whether they were able to correctly rank the words and colors used in the outlook. Whether or not participants correctly ordered the words and colors became our two dependent variables, which we compared to our list of demographic variables through two logistic models, one for word ranking and one for color ranking. These models estimate

TABLE 2. Survey question wording and answer choices.

Question group	Question wording
SPC word ranking	The National Weather Service Storm Prediction Center uses the following phrases to describe the risk of severe thunderstorms and tornadoes. We want to know what these phrases mean to you. Can you rank them from one (lowest risk) to five (highest risk)? (*Words are Marginal, Slight, Enhanced, Moderate, and High, randomly assigned across the five ranks.*)
SPC color ranking	The Storm Prediction Center also uses colors to describe the risk of severe thunderstorms and tornadoes. We want to know what these colors mean to you. Can you rank these colors from one (lowest risk) to five (highest risk)? (*Colors are Green, Yellow, Orange, Red, and Magenta, randomly assigned across the five ranks.*)
Age	How old are you? (*numeric response*)
Gender	Are you male or female? (*Choice of Male or Female*)
Race	Which of the following best describes your race? (*Choice of White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Two or more races, or Some other race*)
State	Please select the state or district where your primary residence is located (*drop down list*)
Education	What is the highest level of education you have COMPLETED? (*Choose from Less than high school, High school/GED, Vocational or Technical Training, Some College; NO degree, 2-year College/Associate's degree, Bachelor's Degree, Master's Degree, or PhD/JD (law)*)
Follow the weather	I follow the weather very closely. (*5-point Likert scale, Strongly Disagree to Strongly Agree*)
Tornado watch/warning comprehension	This alert is issued when severe thunderstorms and tornadoes are possible in and near the area. It does not mean that they will occur. It only means they are possible. (*Select Tornado WATCH (correct), Tornado WARNING, or Do not know*)
	This alert is used when a tornado is imminent. When this alert is issued, seek safe shelter immediately. (*Select Tornado WATCH, Tornado WARNING (correct), or Do not know*)
Time spent	This is a measure of the time in seconds that participants spent on the page with the two SPC ranking questions. Times were broken down by quartile, with a value of 0 given to participants below the first quartile, up to a value of 3 for participants above the third quartile.

the probability that a given participant correctly ranked the outlook words or colors, based on their individual characteristics. The equation for these models included nine independent variables: age, gender, race, education, numeracy, NWS region, following the weather, time spent on the question page, and objective tornado watch and warning understanding. The models for color and word ranking were identical save for their dependent variables. We also tested for interactions between numeracy and time spent on the question page and the other independent variables, but only found one statistically significant relationship between time spent on the page and region, and thus chose to not include these versions of the models in this discussion.

4. Results

a. Public interpretation of the convective outlook

First, the distributions for participant's preferred words and colors for the five ranking levels and their ranking scores were plotted to compare participant's raw responses. These distributions revealed some potential points of confusion for participants, as well as parts of the SPC scale that work effectively. Figure 4 reveals that participants tended to interchange the positions of the words Marginal and Slight, as well as Moderate and Enhanced, when ordering the words from least to greatest risk. The first four risk words as favored by the participants were thus Slight, Marginal, Moderate, and Enhanced, in that order. Over 60% of participants chose the word High for the fifth word in this sample.

Indeed, the largest group of participants in this survey ordered the SPC category words as ''Slight, Marginal, Moderate, Enhanced, High,'' in contrast to the order used in the official product (Table 3). As in Fig. 4, the most common ranking order participants reported swapped the rank order of Marginal and Slight, as well as Enhanced and Moderate. The official order of the SPC words was the fourth largest group of participants, below groups that interchanged the positions of only Marginal and Slight and only Enhanced and Moderate. High was consistently perceived to be the highest risk word in the different ordering groups, except for the fifth and eighth largest groups.

Participant responses by rank for the SPC colors were closer to the official rank order used in the convective outlook (Fig. 5). Green, yellow, and orange were the most commonly chosen colors for the first three ranks, although 22% and 17% of participants chose magenta for the second and third ranks, respectively. For the fourth rank, red was the third most common choice at only 19% of participants, behind magenta and orange at 32% and 28%, respectively. Red was overwhelmingly chosen by participants as the fifth risk color.

The color rank order results show that a plurality of participants switched the positions of magenta and red in their preferred color order (see Table 4). The official color order used by the SPC was the second most common order in this group of participants. Interestingly, red and green were

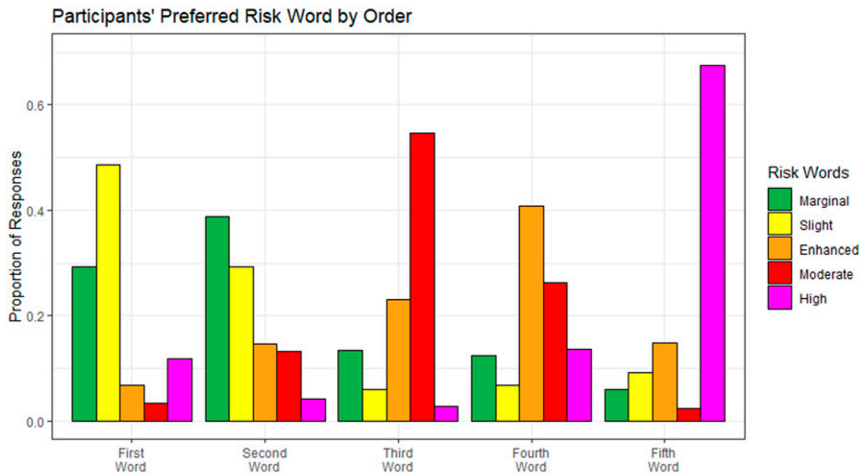

FIG. 4. Distributions of survey participants' ranking of the SPC categorical words.

almost universally chosen as the fifth and first ranked colors by participants, with the position of magenta varying within a more stable order.

b. Demographics versus convective outlook interpretation

A set of independent variables were then used to predict whether participants correctly ranked the word and color scores in a set of logistic regression models (Table 5). The two models revealed no significant relationships between participant's ranking scores and their NWS region of residence. Gender and education only had significant effects on whether a participant correctly ranked the SPC words. However, age, race, tornado watch/warning comprehension, and time spent on the page with the ranking problems were significantly related to whether participants correctly ranked the SPC colors. Numeracy was the only independent variable that had a significant effect on correctly ranking both the words and colors. The word and color models reported pseudo R^2 values of 0.07 and 0.19, respectively, which suggest the independent variables included in these models only explain a small amount of the observed variance in participants' outlook interpretation.

For the word ordering task, the likelihood that participants identified the correct order increased with numeracy and education, and was higher for female participants (see Fig. 6). Participants with lower numeracy (numeracy = 1) had a significantly lower probability of correctly ordering the words [Pr(correct) = 0.06] than those with higher levels of numeracy [numeracy = 7, Pr(correct) = 0.16, see Fig. 6a]. Participants without a high school degree (education = <high school) had a lower probability of correctly ranking the outlook words [Pr(correct) = 0.06] than those with advanced college degrees [education = PhD/JD/MD, Pr(correct) = 0.12, see Fig. 6b], though the difference across educational levels was less than that across numeracy. Finally, female participants (gender = female) were a bit more likely to correctly rank the SPC outlook words [Pr(correct) = 0.08] than males [gender = male, Pr(correct) = 0.05, see Fig. 6c].

There were several more significant relationships between the independent variables and the likelihood that participants correctly ordered the SPC outlook colors (see Fig. 7). Older participants (age = >75) were significantly less likely to correctly order the outlook colors [Pr(correct) = 0.05] than younger participants [age = 18–29, Pr(correct) = 0.18, see Fig. 7a]. Participants with lower numeracy (numeracy = 1), similar to the word ordering task, were much less likely to choose the correct color order [Pr(correct) = 0.06] than more

TABLE 3. The number of participants that ranked the order of the SPC words the same way, for the eight ranking groups with n greater than 90. The words participants selected for each position are labeled, as well as the number n of participants for each group. The official SPC rank order is italicized.

First word	Second word	Third word	Fourth word	Fifth word	n	Percent of participants (%)
Slight	Marginal	Moderate	Enhanced	High	505	16.80
Marginal	Slight	Moderate	Enhanced	High	369	12.28
Slight	Marginal	Enhanced	Moderate	High	224	7.45
Marginal	*Slight*	*Enhanced*	*Moderate*	*High*	*205*	*6.82*
Slight	Marginal	Moderate	High	Enhanced	158	5.26
Slight	Moderate	Marginal	Enhanced	High	107	3.56
Slight	Enhanced	Marginal	Moderate	High	105	3.49
Marginal	Slight	Moderate	High	Enhanced	99	3.29

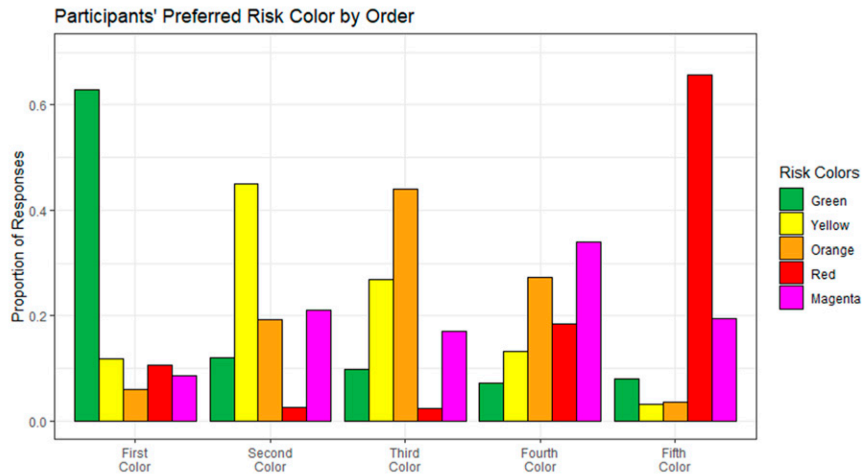

FIG. 5. Distributions of survey participants' ranking of the SPC categorical colors.

numerate participants [numeracy = 7, Pr(correct) = 0.30, see Fig. 7b]. However, we also identified that participants were more likely to correctly rank the outlook colors when they spent more time on the question page [time spent = >61.25, Pr(correct) = 0.13] than those who spent less time on the page [time spent = <29, Pr(correct) = 0.08, see Fig. 7c]. Race was also related to participants' ability to correctly rank the outlook colors, with White participants (race = White) more likely to correctly rank the outlook colors [Pr(correct) = 0.10] than Black participants [race = Black, Pr(correct) = 0.03] and participants of other racial groups [race = Other, Pr(correct) = 0.06, see Fig. 7d]. Finally, participants who could correctly define the difference between a tornado watch and a tornado warning (watch/warning understanding = correct) were slightly more likely to also correctly rank the SPC colors [Pr(correct) = 0.10] than those who could not [watch/warning understanding = incorrect, Pr(correct) = 0.07, see Fig. 7e].

5. Conclusions

Few studies have investigated directly how members of the public interpret the words and colors used in SPC convective outlook, as the product was not originally designed for public use. Nevertheless, the increasing visibility of the outlook on television and social media necessitates investigation into how well the public understands the information it provides. Our results show that the outlook words are not easily ranked in order of risk level by members of the public, although the outlook colors are overall more accurately ordered. Individuals who are more numerate were found to be significantly better able to accurately rank the SPC outlook words and colors. Other factors, such as gender for the word ranking, and age, race, time spent on the questions, and understanding of tornado watches and warnings for the color ranking, were also found to impact individual ranking tasks. These results suggest that the words used in the outlook may lead to wide variations in severe weather risk interpretation by members of the public, particularly for the low and medium probability events covered by Marginal, Slight, Enhanced, and Moderate. For example, people told they are at a Moderate risk for severe weather may underweight the threat posed to them on that day, versus on an Enhanced risk day where they could overweight the severe weather threat. This could lead to a false alarm effect and loss of trust in weather forecast providers, especially among those who are less numerate, when severe weather is more or less serious than expected. Worse, these misunderstandings could increase the vulnerability of people in harm's way if they do not take protective actions they might have otherwise taken if the severe weather risk was correctly communicated to them.

TABLE 4. The number of participants that ranked the order of the SPC colors the same way, for the eight ranking groups with *n* greater than 90. The words participants selected for each position are labeled, as well as the number *n* of participants for each group. The official SPC rank order is italicized.

First color	Second color	Third color	Fourth color	Fifth color	*n*	Percent of participants (%)
Green	Yellow	Orange	Magenta	Red	477	15.87
Green	*Yellow*	*Orange*	*Red*	*Magenta*	*352*	*11.71*
Green	Magenta	Yellow	Orange	Red	277	9.22
Green	Yellow	Magenta	Orange	Red	202	6.72
Green	Orange	Yellow	Magenta	Red	154	5.12
Green	Magenta	Orange	Yellow	Red	97	3.23

TABLE 5. The statistical output, converted into marginal effects (change in percentage correct for each unit increase in the independent variable) and Δp values (change in percentage correct from lowest to highest level of each independent variable), from the logistic regression models that compared demographics to SPC word and color scores (e.g., males are 3.2% less likely to correctly rank the SPC words than females). Standard errors are reported in parentheses where one asterisk (*) indicates $p < 0.05$, two asterisks (**) $p < 0.01$, and three asterisks (***) $p < 0.001$. The McKelvey–Zavoina pseudo R^2 scores for each model are included in the bottom row of the table.

	Word score model		Color score model	
Independent variables	Marginal effects	Δp	Marginal effects	Δp
Age group	0.000 (0.004)	0.001	−0.032*** (0.005)	−0.132
Male (vs female)	−0.032*** (0.009)	−0.032	0.009 (0.011)	0.009
Black (vs White)	−0.011 (0.014)	−0.011	−0.069*** (0.012)	−0.069
Other race (vs White)	−0.015 (0.014)	−0.015	−0.045* (0.015)	−0.045
Education	0.007** (0.003)	0.065	−0.001 (0.003)	−0.004
Numeracy	0.011*** (0.003)	0.103	0.027*** (0.003)	0.243
Eastern region (vs central region)	−0.013 (0.011)	−0.013	−0.014 (0.013)	−0.014
Southern region (vs central region)	−0.009 (0.012)	−0.009	0.020 (0.015)	0.020
Western region (vs central region)	−0.008 (0.012)	−0.008	−0.018 (0.015)	−0.018
Follow the weather	0.000 (0.005)	0.001	0.007 (0.005)	0.028
Tornado watch/warning comprehension	0.016 (0.011)	0.016	0.039** (0.012)	0.039
Time spent on questions	−0.001 (0.004)	−0.005	0.017** (0.005)	0.056
McKelvey–Zavoina pseudo R^2	0.077		0.190	

Participants had the greatest difficulty ordering the first four words in the outlook, consistently swapping the positions of marginal and slight, as well as enhanced and moderate. The color magenta was also difficult for participants to place, though the first three colors were generally listed in the correct order, and red was consistently placed as the highest risk color. Overall, the colors appeared to be more in line with user expectations of risk communication, which would likely be aided

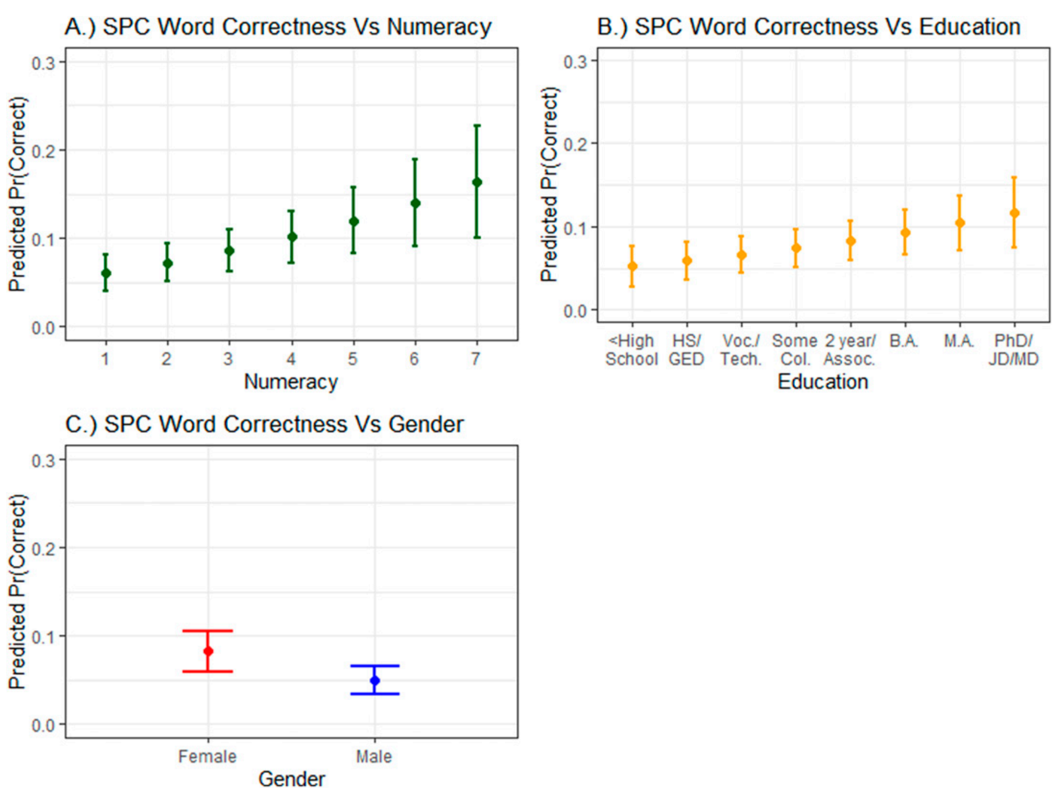

FIG. 6. Graphs display the likelihood that a participant with a given trait will correctly order the words used in the SPC outlook, based on (a) numeracy, (b) education, and (c) gender. Points display the simulated mean probability, all other variables held equal, while the error bars display the 95% confidence interval around that estimate based on the standard error.

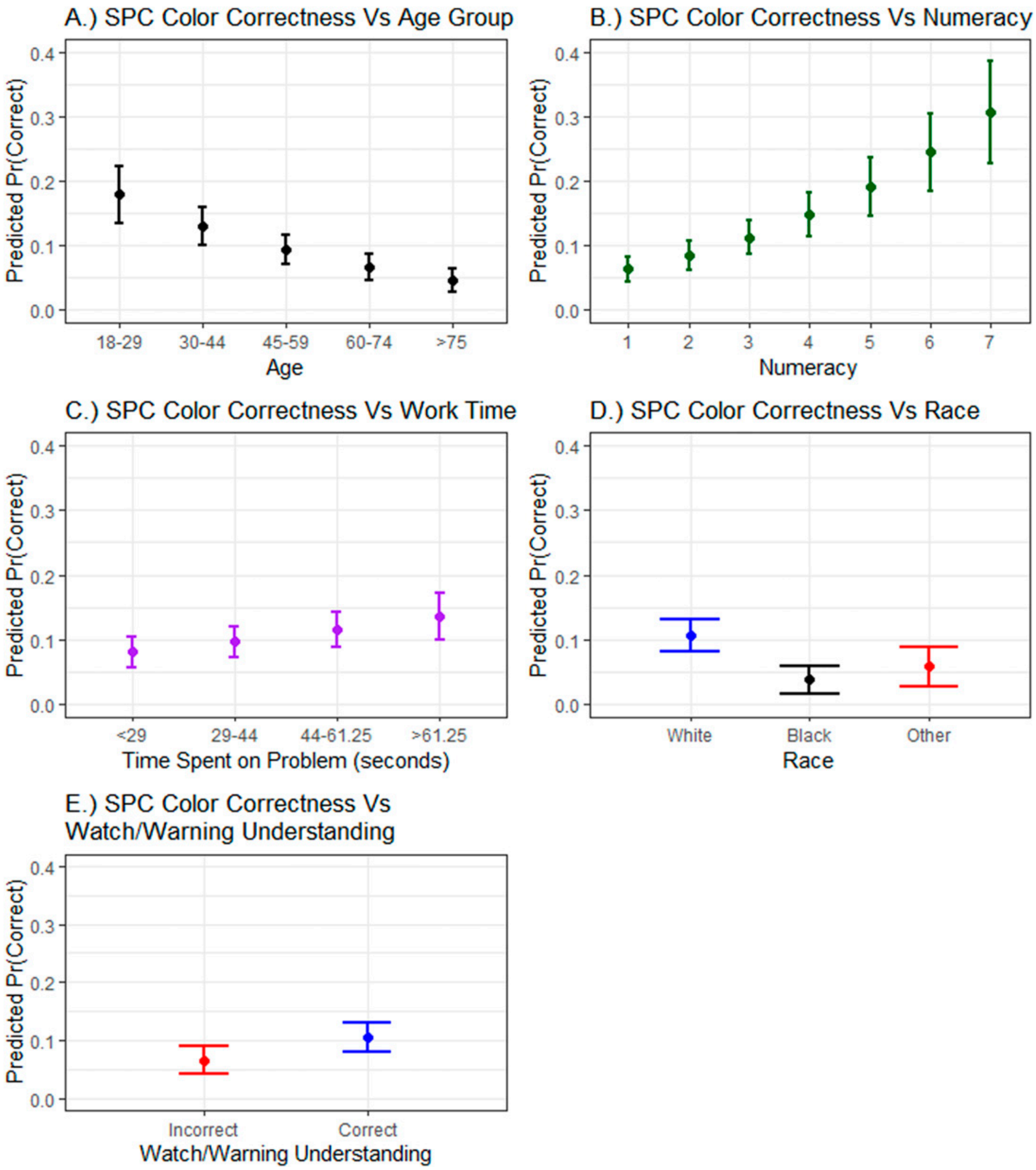

FIG. 7. Graphs display the likelihood that a participant of a given demographic will correctly order the words used in the SPC outlook, based on (a) age, (b) numeracy, (c) time spent on the questions, (d) racial group, and (e) tornado watch vs warning understanding. Points display the simulated mean probability, all other variables held equal, while the error bars display the 95% confidence interval around that estimate based on the standard error.

by the bull's-eye pattern in which the colors are presented in the visual outlook product. The outlook words, however, can be encountered outside of the visual context of the product image. This is especially concerning due to the differences between how participants ranked the risk communicated by the words and the ranks used in the product. Future work should seek to identify whether other versions of the outlook words and/or additions to the current set of words (such as risk levels or probability numbers) can be better tuned to the risk that individuals feel those words communicate. Additionally, the large swings we identified in outlook ranking ability across numeracy could suggest social vulnerabilities in the nation's severe weather messaging paradigm that should be addressed. Further vulnerabilities appear to exist across racial groups, as non-White participants on average struggled more to interpret the SPC risk colors, something also seen in English second language and non-English speaking communities in the United States (Trujillo-Falcón et al. 2021). Future work testing NWS products like the SPC outlook should seek to investigate disparities in comprehension across individuals with different levels of numeracy, and across a wider range of racial and non-English speaking groups, as these differences may be

TABLE A1. Questions asked as part of the BNT-S numeracy test.

Question and wording	Percent of respondents correct
Imagine that we flip a fair coin 1000 times. What is your best guess about how many times the coin would come up heads in 1000 flips? (*Verbatim, Answer = 500.*)	52.93%
In the BIG BUCKS LOTTERY, the chance of winning a $10 prize is 1%. What is your best guess about how many people would win a $10 prize if 1000 people each buy a single ticket to BIG BUCKS? (*Verbatim, Answer = 10.*)	44.33%
In ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car? (*Verbatim, Answer = 0.1*)	14.73%
Out of 1000 people in a small town 500 are members of a choir. Out of these 500 members in a choir 100 are men. Out of the 500 inhabitants that are not in a choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability as a percent. (*Verbatim, Answer = 25*)	11.03%
Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? (*Verbatim, Answer = 30*)	24.97%
Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of 70 throws how many times would the die show the number 6? (*Verbatim, Answer = 20*)	12.93%
In a forest, 20% of the mushrooms are red, 50% are brown, and 30% are white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? Please indicate the probability as a percent. (*Verbatim, Answer = 50*)	7.27%

systematic across other NWS products as well. This research could also help guide more targeted outreach efforts with groups less familiar with the NWS product suite. Finally, proposed changes should also be evaluated with core partners, such as emergency managers and broadcast meteorologists, who are familiar with the current system, to ensure that any changes to the outlook design do not hinder their work.

Future studies could also work to address some of the limitations of this study. Though the sampling strategy for this study was stratified to match U.S. Census estimates, the method through which questions are asked can lead to errors in the data collected. For example, the ordering task did not include a ''do not know'' option and randomly assigned each word and color to a random position in the administered survey, meaning some erroneous responses created by random generation instead of participant input cannot be identified in the data. These nonresponses may have been more of an issue for the color ranking task, which participants were asked to complete after the word ranking task and was positively related to how long participants spent on the survey page that contained both questions. We also did not test participants' interpretation of the full convective outlook, as participants were instead asked to rank the words and colors in separate survey questions with no accompanying visual aid. Participant interpretation of the words and colors in the context of the outlook graphic may differ due to these two elements being combined and displayed in a way that allows individuals to infer the correct risk ordering. Finally, our models for outlook word and color interpretation were only able to explain 7% and 19% of the observed variance in interpretation ability, suggesting future work could identify further variables and develop more complex models that better explain how members of the public interpret the SPC outlook.

Based on the findings of this study, however, it is clear that work must be done to improve SPC outlook comprehension with the general public if the product is to successfully communicate risk to nonexpert users. One way to do this would be for the weather enterprise to increase its emphasis on educating students in the K–12 system, through meteorologists directly interacting with students and teaching about weather safety and information sources they can use, including the SPC outlook. A complimentary method would be to address the design of the SPC outlook and develop a more intuitive method of risk communication that requires less in-depth education for users from core partners to the public to interpret. Regardless of how user interpretation is improved, the potential for this product as a priming tool to ready members of the public to receive tornado warnings, or to prepare their emergency supplies, plans, and shelters, can only be fully realized if the convective outlook is readily understood by nonexpert users. As the SPC is currently working to develop more precise forecasts of severe storm intensity in the convective outlook, work should also continue to be done to understand how the final outlook product could be better formatted for public understanding and use.

Acknowledgments. The authors thank their team and coworkers at the Center for Risk and Crisis Management for all their support and assistance in writing this paper. We also thank Dr. Patrick Marsh and Dr. Israel Jirak at the Storm Prediction Center for their help in identifying information about the history of the convective outlook product and for helping proofread the paper. Data collection for this project was funded by the OU Office of the Vice President for Research. Data analysis was funded by National Oceanic and Atmospheric Administration Project OAR-USWRP-R2O, ''FACETs Probability of What? Understanding and Conveying Uncertainty through Probabilistic Hazard Services,'' and National Oceanic and Atmospheric Administration Project NA18OAR4590376, ''Communicating Forecast Uncertainty and Probabilistic

Information: Experimenting with Social Observation Data in the Hazardous Weather Testbed."

Data availability statement. Data are available at https://dataverse.harvard.edu/dataverse/wxsurvey.

APPENDIX

Table of Questions Used to Measure Numeracy

Table A1 contains the full list of questions that can be asked as a part of the BNT-S adaptive numeracy test.

REFERENCES

Allan, J. N., 2018: Numeracy vs. intelligence: A model of the relationship between cognitive abilities and decision making. M.S. thesis, Dept. of Psychology, The University of Oklahoma, 88 pp.

——, J. T. Ripberger, V. T. Ybarra, and E. T. Cokely, 2017: The Oklahoma warning awareness scale: A psychometric analysis of a brief self-report survey instrument. *Proc. Human Factors and Ergonomics Society 2017 Annual Meeting*, Austin, TX, Human Factors, 1203–1207, https://doi.org/10.1177/1541931213601783.

——, W. W. Wehde, J. T. Ripberger, C. Silva, and H. Jenkins-Smith, 2019: The geographic distribution of extreme weather and climate risk perceptions in the United States. *Seventh Symp. on Building a Weather-Ready Nation*, Phoenix, AZ, Amer. Meteor. Soc., 676, https://ams.confex.com/ams/2019Annual/meetingapp.cgi/Paper/351780.

Bryant, B., M. Holiner, R. Kroot, K. Sherman-Morris, W. B. Smylie, L. Stryjewski, M. Thomas, and C. I. Williams, 2014: Usage of color scales on radar maps. *J. Oper. Meteor.*, **2**, 169–179, http://dx.doi.org/10.15191/nwajom.2014.0214.

Cappucci, M., 2020: The National Weather Service issues highly accurate thunderstorm forecasts. The public doesn't understand them. *The Washington Post*, accessed 28 July 2020, https://www.washingtonpost.com/weather/2020/06/10/storm-prediction-center-risk-categories/.

Cho, J., E. T. Cokely, M. Ramasubramanian, J. N. Allan, A. Feltz, and R. Garcia-Retamero, 2021: Risk literacy promotes representative understanding: Numerate people are less biased, more knowledgeable, and more concerned about climate change. *Res. Square*, https://doi.org/10.21203/rs.3.rs-420681/v1.

Cokely, E. T., M. Galesic, E. Schulz, S. Ghazal, and R. Garcia-Retamero, 2012: Measuring risk literacy: The Berlin Numeracy Test. *Judgment Decision Making*, **7**, 25–47.

——, A. Feltz, S. Ghazal, J. N. Allan, D. Petrova, and R. Garcia-Retamero, 2018: Skilled decision theory: From intelligence to numeracy and expertise. *The Cambridge Handbook of Expertise and Expert Performance.* K. A. Ericsson et al., Eds., Cambridge University Press, 476–505.

Corfidi, S. F., 1999: The birth and early years of the Storm Prediction Center. *Wea. Forecasting*, **14**, 507–525, https://doi.org/10.1175/1520-0434(1999)014<0507:TBAEYO>2.0.CO;2.

Edwards, R., and F. Ostby, 2015: Time line of SELS and SPC. Storm Prediction Center, accessed 28 February 2020, https://www.spc.noaa.gov/history/timeline.html.

Ernst, S., D. LaDue, and A. Gerard, 2018: Understanding emergency manager forecast use in severe weather events. *J. Oper. Meteor.*, **6**, 95–105, https://doi.org/10.15191/nwajom.2018.0609.

Gerst, M. D., and Coauthors, 2020: Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks. *Wea. Climate Soc.*, **12**, 117–133, https://doi.org/10.1175/WCAS-D-18-0094.1.

Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, https://doi.org/10.1175/WAF-D-17-0104.1.

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting*, **27**, 1580–1585, https://doi.org/10.1175/WAF-D-12-00061.1.

——, and ——, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, https://doi.org/10.1175/WAF-D-13-00132.1.

——, and ——, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. *Wea. Forecasting*, **32**, 1321–1328, https://doi.org/10.1175/WAF-D-16-0170.1.

——, ——, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Jauernic, S. T., and M. S. Van Den Broeke, 2017: Perceptions of tornadoes, tornado risk, and tornado safety actions and their effects on warning response among Nebraska undergraduates. *Nat. Hazards*, **80**, 329–350, https://doi.org/10.1007/s11069-015-1970-9.

Lipkus, I. M., and J. G. Hollands, 1999: The visual communication of risk. *J. Natl. Cancer Inst.*, **1999**, 149–163, https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191.

——, G. Samsa, and B. K. Rimer, 2001: General performance on a numeracy scale among highly educated samples. *Med. Decis. Making*, **21**, 37–44, https://doi.org/10.1177/0272989X0102100105.

Mason, J. B., and J. C. Senkbeil, 2015: A tornado watch scale to improve public response. *Wea. Climate Soc.*, **7**, 146–158, https://doi.org/10.1175/WCAS-D-14-00035.1.

Morrow, B. H., J. K. Lazo, J. Rohme, and J. Feyen, 2015: Improving storm surge risk communication: Stakeholder perspectives. *Bull. Amer. Meteor. Soc.*, **96**, 35–48, https://doi.org/10.1175/BAMS-D-13-00197.1.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

NCDC, 2020: Billion-dollar weather and climate disasters: Time series. NOAA, accessed 28 July 2020, https://www.ncdc.noaa.gov/billions/time-series/US.

NOAA, 2011: NWS Central Region service assessment: Joplin, Missouri, tornado—May 22, 2011. National Weather Service, 41 pp., https://www.weather.gov/media/publications/assessments/Joplin_tornado.pdf.

Pietrycha, A. E., and M. A. Fox, 2004: Effective use of various communication methods during a severe convective outbreak. *NWA Digest*, **28**, 59–64.

Powell, S. W., and H. D. O'Hair, 2008: Communicating weather information to the public: People's reactions and understandings of weather information and terminology. *Third Symp. on Policy and Socio-Economic Research*, New Orleans, LA, Amer. Meteor. Soc., P1.3, https://ams.confex.com/ams/88Annual/webprogram/Paper132939.html.

Ripberger, J. T., M. J. Krocak, W. W. Wehde, J. N. Allan, C. Silva, and H. Jenkins-Smth, 2019: Measuring tornado warning reception, comprehension, and response in the United States. *Wea. Climate Soc.*, **11**, 863–880, https://doi.org/10.1175/WCAS-D-19-0015.1.

——, C. L. Silva, H. C. Jenkins-Smith, J. Allan, M. Krocak, W. Wehde, and S. Ernst, 2020: Exploring community differences in tornado warning reception, comprehension, and response across the United States. *Bull. Amer. Meteor. Soc.*, **101**, E936–E948, https://doi.org/10.1175/BAMS-D-19-0064.1.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Schultz, D. M., E. C. Gruntfest, M. H. Hayden, C. C. Benight, S. Drobot, and L. R. Barnes, 2010: Decision making by Austin, Texas, residents in hypothetical tornado scenarios. *Wea. Climate Soc.*, **2**, 249–254, https://doi.org/10.1175/2010WCAS1067.1.

Schwartz, L. M., S. Woloshin, W. C. Black, and H. G. Welch, 1997: The role of numeracy in understanding the benefit of screening mammography. *Ann. Intern. Med.*, **127**, 966–972, https://doi.org/10.7326/0003-4819-127-11-199712010-00003.

Silva, C., J. T. Ripberger, H. Jenkins-Smith, M. Krocak, S. Ernst, and A. Bell, 2019: Establishing a baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. Reference Rep., 33 pp.

Stough, S., E. M. Leitman, J. L. Peters, and J. Correia Jr., 2012: The role of the Storm Prediction Center products in decision making leading up to severe weather events. NOAA, 18 pp., https://www.spc.noaa.gov/publications/leitman/stough.pdf.

Trujillo-Falcón, J. E., O. Bermúdez, K. Negrón-Hernández, J. Lipski, E. Leitman, and K. Berry, 2021: Hazardous weather communication en Español: Challenges, current resources, and future practices. *Bull. Amer. Meteor. Soc.*, **102**, E765–E773, https://doi.org/10.1175/BAMS-D-20-0249.1.

WPC, 2020: Excessive rainfall forecasts. Weather Prediction Center, accessed 28 February 2020, https://www.wpc.ncep.noaa.gov/qpf/excess_rain.shtml.

Williams, C. A., A. J. Grundstein, and J. So, 2020: Should severe weather graphics wear uniforms? Understanding the effects of inconsistent convective outlook graphics on members of the public. *15th Symp. on Societal Applications: Policy, Research, and Practice*, Boston, MA, Amer. Meteor. Soc., 11A.1, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/365011.