Measuring Tornado Warning Reception, Comprehension, and Response in the United States

JOSEPH T. RIPBERGER

University of Oklahoma Center for Risk and Crisis Management, and National Institute for Risk and Resilience, Norman, Oklahoma

MAKENZIE J. KROCAK

University of Oklahoma Center for Risk and Crisis Management, and National Institute for Risk and Resilience, and Cooperative Institute for Mesoscale Meteorological Studies, and NOAA/Office of Oceanic and Atmospheric Research (OAR)/National Severe Storms Laboratory, Norman, Oklahoma

WESLEY W. WEHDE,[a] JINAN N. ALLAN, CAROL SILVA, AND HANK JENKINS-SMITH

University of Oklahoma Center for Risk and Crisis Management, and National Institute for Risk and Resilience, Norman, Oklahoma

(Manuscript received 25 January 2019, in final form 26 July 2019)

ABSTRACT

Social criteria are important to achieving the mission of the National Weather Service. Accordingly, researchers and administrators at the NWS increasingly recognize a need to supplement verification statistics with complementary data about society in performance management and evaluation. This will require significant development of new capacities to both conceptualize relevant criteria and measure them using consistent, transparent, replicable, and reliable measures that permit generalizable inference to populations of interest. In this study, we contribute to this development by suggesting three criteria that require measurement (forecast and warning reception, comprehension, and response) and demonstrating a methodology that allows us to measure these concepts in a single information domain—tornado warnings. The methodology we employ improves upon previous research in multiple ways. It provides a more generalizable approach to measurement using a temporally consistent set of survey questions that are applicable across the United States; it relies on a more robust set of psychometric tests to analytically demonstrate the reliability of the measures; and it is more transparent and replicable than previous research because the data and methods (source code) are publicly available. In addition to describing and assessing the reliability of the measures, we explore the sensitivity of the measures to geographic and demographic variation to identify significant differences that require attention in measurement. We close by discussing the implications of this study and the next steps toward development and use of social criteria in performance management and evaluation.

1. Introduction

The mission of the National Weather Service (NWS) is to provide weather, water, and climate data, forecasts, and warnings for the protection of life and property and the enhancement of the national economy. Currently, the NWS uses forecast and warning verification statistics (such as probability of detection, false alarm ratio, and warning lead time) to measure the extent to which they are achieving this mission. This strategy presumes that increasing the accuracy and timeliness of forecasts and warnings will reduce loss of life and damage to property due to extreme weather and climate events. Improvements to forecasts and warnings, while necessary, are not sufficient to generate this outcome. Rather, there is a set of social criteria that also must be met. Information users (members of the public, emergency managers, etc.) must receive, comprehend, and respond to the forecasts and warnings that the NWS issues (Drabek 1986; Mileti and Sorensen 1990; Lindell and Perry 2012).

[a] Current affiliation: Department of Political Science, International Affairs, and Public Administration, East Tennessee State University, Johnson City, Tennessee.

Corresponding author: Joseph T. Ripberger, jtr@ou.edu

DOI: 10.1175/WCAS-D-19-0015.1

If people do not receive forecasts and warnings, they cannot use them to make protective action decisions; if people do not understand the information in forecasts and warnings, they cannot use them to make risk-aware choices about how to protect themselves; and if people do not engage in some sort of protective action *in response to the forecasts and warnings,* these products will not reduce loss of life and property damage.

These and other social criteria, such as community preparation and resilience, are important to achieving the NWS mission. As such, researchers and administrators at the NWS increasingly recognize a need to supplement verification statistics with complementary data about society in performance management and evaluation (NOAA 2015; National Academies of Sciences, Engineering, and Medicine 2018). This will require significant development of new capacity to both conceptualize relevant criteria and measure them using consistent, transparent, replicable, and reliable measures that permit generalizable inference to populations of interest. Here, we hope to contribute to this development by suggesting three criteria that require measurement—forecast and warning reception, comprehension, and response. In addition, we demonstrate an approach to measure these concepts in a single information domain—tornado warnings. To do this, we first review previous attempts to measure tornado warning reception, comprehension, and response, focusing on data collection methodologies, concept operationalization, and findings. Second, we introduce our methodology and measures and explain how they relate to and depart from previous research. Third, we assess the reliability of our measures using a variety of psychometric statistics. Fourth, we explore variation in the measures by using them to identify geographic and demographic differences in tornado warning reception, comprehension, and response. We close by discussing the strengths and limitations of the measures and outlining the steps that will be necessary to evaluate and improve NWS operations over time and across jurisdictions.

2. Previous research

a. Tornado warning reception

Previous research has sought to explain when, how, and why people take protective actions in response to information about hazards and disasters (e.g., Drabek 1986; Mileti and Sorensen 1990; Lindell and Perry 2012). These models collectively reject the ''stimulus-response'' idea that information about risk reflexively causes people to take protective action. Instead they embrace the notion that protective action requires a complex multistage social process that begins with information reception. Simply put, people must receive and pay attention to the information before they can do something with it. Recognizing this, a variety of studies measure public reception of tornado warnings that are issued by the NWS. Many of these studies use postevent data collection methodologies to document warning reception in geographically specific populations following a significant tornado (e.g., Hammer and Schmidlin 2002; Mitchem 2003; Paul et al. 2015) or a tornado warning (e.g., Godfrey et al. 2011). For example, Hammer and Schmidlin (2002) measure warning reception along the path of damage done by the Bridge Creek–Moore tornado on 3 May 1999. Given the relatively small populations that these studies target, many of them collect data in the field, by way of in-person interviews (e.g., Mitchem 2003). Others use surveys, primarily relying on convenience sampling to recruit participants (e.g., Comstock and Mallonee 2005; Biddle 2007; Sherman-Morris 2010; Paul and Stimers 2012; Jauernic and Van Den Broeke 2016). While valuable, it is difficult to make inferences about the general population from convenience samples. To overcome this, a few studies employ random or representative sampling techniques to ensure that study participants are representative of the characteristics of the target population—those affected, nationwide, or otherwise (e.g., Brown et al. 2002; Godfrey et al. 2011).

Previous studies also operationalize tornado warning reception in different ways. Some measure the concept as a simple dichotomy, asking respondents to indicate if they did or did not receive a warning before the event in question (Mitchem 2003; Godfrey et al. 2011). More commonly, previous studies operationalize reception by asking interview or survey participants to identify warning sources such as sirens, television, or friends and family (Balluz et al. 2000; Brown et al. 2002; Hammer and Schmidlin 2002; Comstock and Mallonee 2005; Biddle 2007; Sherman-Morris 2010; Paul and Stimers 2012; Paul et al. 2015).

Despite the variety and complexity of target populations, methodologies, and operationalizations, previous studies report relatively consistent findings. Between 80% and 90% of people say that they receive relevant tornado warnings (Brown et al. 2002; Hammer and Schmidlin 2002; Comstock and Mallonee 2005; Biddle 2007; Godfrey et al. 2011; Paul et al. 2015). Despite this consistency, there is some evidence that warning reception varies across demographic groups. Mitchem (2003), for example, finds that reception may increase with age, that White individuals are less likely to receive warnings than Black individuals, and that men are less likely than women to receive warnings. Aguirre (1988) observes that reception varies with ethnicity as well, especially among Hispanic populations who rely on Spanish television broadcasts. Although valuable, these studies rely on small samples

of survey respondents in Marion County, Indiana, and Saragosa, Texas, so the results may not generalize to adults across the United States.

b. Tornado warning comprehension

Scholars agree that warning comprehension is necessary to promote protective action decision making (Drabek 1986; Mileti and Sorensen 1990; Lindell and Perry 2012). People must understand the contents of a message if they are going to use it to assess risk and rationally evaluate response options. Although it is critically important, relatively few studies systematically measure tornado warning comprehension. Studies that use interviews or surveys to collect data in geographically specific populations usually rely on convenience sampling to recruit study participants. In a few studies, data collection follows and references a specific event (e.g., Chaney and Weaver 2010). Others measure comprehension more generally, without reference to a specific event (e.g., Powell and O'Hair 2008).

This previous work also conceptualizes warning comprehension in a relatively narrow way, focusing almost exclusively on objective ability to distinguish between tornado watches and warnings. Studies assess this ability with different types of test questions. Some use multiple choice questions that ask study participants to select the correct definition given an alert (e.g., Chaney and Weaver 2010; Chaney et al. 2013; Schultz et al. 2010). Other studies use open-ended questions that ask respondents to provide a definition for the different alerts. This strategy eliminates the possibility that subjects will guess the correct response but puts the onus on researchers to decide what constitutes a correct response. Some studies do this with binary (correct/incorrect) judgements (e.g., Powell and O'Hair 2008; Mason and Senkbeil 2015); others use scales to rate correctness (e.g., Sherman-Morris 2010).

This variety in studied populations and testing procedures has made previous assessments of tornado warning comprehension fairly inconsistent. Balluz et al. (2000) presents the most optimistic findings; over 95% of survey participants in Clark and Saline County, Arkansas, understood the difference between tornado watches and warnings. A study in Austin, Texas, finds similarly that 90% of people can identify the difference between a watch and a warning (Schultz et al. 2010); however, a study in Oklahoma, Texas, and California indicates that 58% of people can make this distinction (Powell and O'Hair 2008); and a study in Alabama finds that only 47% of people can (Mason and Senkbeil 2015). Much of this variation likely stems from differences in geography and survey sampling. It is possible that some of this variation stems from demographic differences across the samples. Comprehension appears to increase with age (Powell and O'Hair 2008; Sherman-Morris 2010) and White individuals are more likely to know the difference between tornado watches and warnings than non-White individuals (Powell and O'Hair 2008). Additional variation is likely attributable to differences in geography and corresponding exposure to watches and warnings. Powell and O'Hair (2008) provides preliminary support for this conjecture, finding that 68% and 70% of Oklahomans and Texans, respectively, can differentiate between watches and warnings, whereas only 50% of Californians are able to do so.

c. Tornado warning response

Following information reception and comprehension, most scholars agree that risk information will only enhance the protection of life and property if it leads people to engage in some form of protective action. Because of this obvious and important connection, virtually all previous studies of when, how, and why people take protective action in response to tornado warnings measure response in some way. Many use postevent interviews and surveys to measure response in specific locations in the immediate aftermath of significant events (e.g., Schmidlin et al. 2009; Miran et al. 2018). Others measure response more generally without reference to specific events (e.g., Klockow 2013; Casteel 2018). Most previous studies use convenience sampling to recruit study participants, but there are a few that employ more generalizable sampling techniques to make sure that study participants match the characteristics of the target population (e.g., Klockow 2013; Ripberger et al. 2015a,b).

Researchers operationalize tornado warning response in many different ways. Often, postevent studies use simple interviews or survey questions that ask study participants if they took protective action when they got the tornado warning and/or what actions they took (e.g., Silver and Andrey 2014; Miran et al. 2018). Studies that do not follow or reference specific events typically operationalize response by asking study participants to indicate future response intentions, with questions asking some variant of, "What do you plan to do the next time you receive a tornado warning?" (e.g., Lindell et al. 2016; Ripberger et al. 2015a,b). It can be challenging for some participants to judge how they, personally, will act in response to future warnings because they know that their actions will be highly dependent on the context of the warning—where they are they, who they are with, what time it is, etc. To address this challenge, some studies measure warning response using "micro decision-making environments" that remove the participant from the scenario by asking them to assume the role of a person who is responsible for making protective action decisions for a business or group of people (e.g., Klockow 2013;

Casteel 2016; Casteel 2018). As with judgments about warning comprehension, measures of warning response that allow study participants to indicate the type of action they took in the past, plan to take in the future, or advise taking in a micro decision-making environments puts the onus on researchers to decide which actions constitute a protective response and which actions do not. This can be easy for some actions (e.g., sheltering in place) and difficult for others (e.g., monitoring the situation).

Postevent studies of tornado warning response report a wide range of findings. For instance, Miran et al. (2018) report that 58%, 79%, and 43% of study participants took protective action in response to three different tornado warnings in Oklahoma City; Chaney et al. (2013) find that 77% of DeKalb County participants took action in response to warnings in Alabama; and Balluz et al. (2000) note that only 45% took action in response to warnings in Clark and Saline County, Arkansas. Studies that measure response by assessing future intentions are a bit more consistent and optimistic, finding that somewhere between 75% and 90% of study participants plan to act the next time they receive a tornado warning (e.g., Schultz et al. 2010; Lindell et al. 2016; Ripberger et al. 2015a,b). Many of these studies indicate that demographic differences impact warning response. For example, responsiveness appears to increase and then decrease with age (e.g., Senkbeil et al. 2012; Chaney et al. 2013), increase with education (e.g., Balluz et al. 2000; Schmidlin et al. 2009), and vary by gender, with men being less responsive than women (e.g., Comstock and Mallonee 2005; Sherman-Morris 2010; Ripberger et al. 2015a; Robinson et al. 2019).

In sum, previous research on tornado warning reception, comprehension, and response is very diverse. Studies operationalize the concepts in different ways, rely on different data collection methodologies, and focus on different geographic areas. This diversity is imperative to scientific inquiry as it allows us to identify the boundaries of theory by measuring and comparing complex concepts in diverse settings using multiple sources of information. Unfortunately, diversity in measurement can complicate the task of performance management and evaluation. In this context, consistency in measurement can be extremely important as it facilitates systematic comparison of concepts over time and across geographic regions. These comparisons are imperative to the protection of life and property because they provide direct feedback to NWS forecasters and officials about the policies and risk communication practices that improve warning reception, comprehension, and response.

3. Research methodology and operationalization

a. Data collection

In addition to consistency, performance management and evaluation requires transparent, replicable, and reliable measures that permit generalizable inference to populations of interest. There are multiple methodologies that might accomplish these goals. For example, meta-analytic methodologies might combine findings from multiple "small" studies that measure tornado warning reception, comprehension, and response in different segments of the population to produce aggregate population estimates that account for uncertainty in previous findings (e.g., Huang et al. 2016). While extremely valuable, meta-analysis requires a relatively large number of small studies and some consistency in measurement across the studies. As we note above, the body of scholarship on tornado warning reception, comprehensions, and response does not (yet) meet these conditions. There are relatively few studies that systematically measure these concepts (especially warning reception and comprehension) and the measures are inconsistent across the studies. Given these limitations, we strive to achieve these goals through the use of consistent large-scale surveys that are geographically and demographically representative of the U.S. population. We call this effort the Severe Weather and Society Survey (WX Survey).

The WX Survey, which began in 2017 (WX17), is a yearly survey of U.S. adults that includes two types of questions: consistent (recurring) questions that measure forecast and warning reception, comprehension, and response; and rotating (one-time) questions and experiments that address important topics in the weather community, such as the impact of uncertainty and probabilistic information on risk judgements and protective action decision making.[1] To facilitate transparency and enhance replicability, the WX Survey team publishes an open access report that presents an overview of the methodology we employ in each survey, the weighting strategy we use to ensure generalizability, and a reproduction of the survey instrument with means and frequencies for each of the questions (Silva et al. 2017; Silva et al. 2018). In addition, the survey data and R code necessary to replicate the measures we create and analysis we conduct are available in a public repository (https://github.com/oucrcm).

[1] Early iterations of the WX Survey were implemented in 2012 and 2013 using a different sample frame (counties that are most likely to experience deadly tornadoes) and a slightly different set of questions.

In combination, the WX Survey methodology provides a basis for developing consistent, transparent, and replicable measures of tornado warning reception, comprehension, and response that permit generalizable inference to U.S. population. In the section that follows, we use data from the 2018 survey (WX18) to address the present research question—the reliability of measures. WX18 was fielded in July 2018 using an online questionnaire that was completed by 3000 U.S. adults (age 18+) across the conterminous United States (CONUS) that match the demographic characteristics of the U.S. population. The sample of participants was provided by Qualtrics, a company that maintains a diverse panel of Internet users in the United States who agree to participate in online surveys. Qualtrics, like most Internet sampling companies, uses a quota system to produce representative samples. In WX18, the quotas give us a diverse sample of survey participants that generally represents the geographic and demographic attributes of the target population (see Table 1). Nevertheless, there are a few imbalances that we address with post-stratification survey weights. To calculate these weights, we divide the proportion of the target population that shares the demographic characteristics of each respondent (the population proportion) by the proportion of the sample that shares these characteristics.

TABLE 1. Representativeness of survey participants. Population estimates were obtained from the U.S. Census annual estimates of the resident population by sex, age, race, and Hispanic origin for the United States and states: 1 Apr 2010 to 1 Jul 2017 (PEPASR6H).

Demographic categories	U.S. adult population (%)	Participants (%)
Gender		
Female	51.3	51.3
Male	48.7	48.7
Age		
18 to 24	12.1	12.2
25 to 34	18.0	18.2
35 to 44	16.2	16.4
45 to 54	16.8	16.6
55 to 64	16.7	16.6
65 and up	20.2	20.0
Ethnicity		
Hispanic	16.1	16.3
Non-Hispanic	83.9	83.7
Race		
White	78.2	76.4
Black or African American	12.8	13.6
Asian	5.8	6.4
Other race	3.2	3.5
NWS region		
Eastern	31.7	31.9
Southern	27.0	26.7
Central	20.7	21.2
Western	20.6	20.2

b. Operationalization

In contrast to previous studies that operationalize tornado warning reception with a single question about a specific warning or event, we use multiple questions that gauge reception across warnings/events. As shown in Table 2, the first three items in this set measure reception more generally; the next items measure reception in different situations; and the last items measure reception at different times of the day. We use this set of multiple items for two reasons. First, the use of multiple items reduces measurement error (Cohen et al. 1996). Second, we recognize that tornado warning reception is a complex construct that varies with circumstance. While we cannot capture *all* circumstances that may influence reception, these measures attempt to capture *some* to encourage participants to think about reception in multiple settings and multiple times throughout the day.

We measure tornado warning comprehension in two ways, objectively and subjectively. For the objective measure, we follow previous studies that use test questions to measure comprehension, but the operationalization we employ captures multiple dimensions of the concept. We ask about basic knowledge of the difference between tornado watches and warnings as well as more in-depth questions about average watch and warning lead times and geographic scopes. We posit that these dimensions are significant because comprehension requires that warning recipients know what the risk is, where the risk is, and when to act if they want to reduce loss of life and property. Table 3 lists the survey questions we use to capture these dimensions. We use language from an NWS safety guidance document on tornado alerts to develop and identify correct responses to the items (NWS 2018).

The subjective measure of tornado warning comprehension relies on multiple items. As shown in Table 4, the first three items measure comprehension generally by asking participants if they recognize the difference between *all* types of watches and warnings, *tornado* watches and warnings, and *severe thunderstorm* watches and warnings. We include the item on severe thunderstorm watches and warnings because they often precede or accompany tornado watches and warnings, so comprehension may require an ability to differentiate between the two. The next two items gauge participant understanding of the risk communication tools that forecasters often use when issuing tornado warnings—maps and radar images. The remaining items tap comprehension

TABLE 2. Operationalization of tornado warning reception. For Rec_All to Rec_Soon: 1—Strongly disagree, 2—Disagree, 3—Neither disagree nor agree, 4—Agree, and 5—Strongly agree; for Rec_Sleep to Rec_Evening: 1—Not at all confident, 2—Not very confident, 3—Somewhat confident, 4—Very confident, and 5—Extremely confident.

Survey question	1	2	3	4	5
Please tell us how strongly you agree with the following statements about tornado WARNINGS:					
Rec_All: I receive all tornado warnings that are issued for my area.	6.5%	15.9%	28.1%	32.4%	17.1%
Rec_Most: I receive most tornado warnings that are issued for my area.	4.9%	7.5%	25.8%	42.9%	18.9%
Rec_Soon: I receive tornado warnings as soon as they are issued for my area.	5.1%	12.2%	29.1%	36.9%	16.7%
Sometimes people *miss* tornado WARNINGS because they are doing something that makes it difficult to pay attention to the weather. For example, people often miss tornado warnings when they are sleeping. How confident are you that you would *receive* tornado warnings in the following situations?					
Rec_Sleep: If you are sleeping?	25.7%	32.7%	21.8%	11.5%	8.3%
Rec_Car: If you are in a car?	7.2%	16.4%	33.9%	27.4%	15.2%
Rec_Work: If you are at work or school?	4.5%	10.8%	29.7%	34.6%	20.4%
Rec_Store: If you are at a store?	7.7%	20.2%	33.5%	25.5%	13.1%
Rec_Sm_Group: If you are with a small group of friends or family?	3.7%	12.1%	37.4%	31.3%	15.4%
Rec_Lg_Group: If you are with a large group of friends or family?	4.2%	12.9%	32.7%	32.8%	17.3%
For some people the time of day influences tornado warning reception, understanding, and/or responsiveness. If a tornado WARNING were issued for your area tomorrow at [RANDOM TIME], how confident are you that you would receive the warning?					
Rec_Morning: 1:00 a.m.–9:00 a.m.	11.3%	16.6%	28.7%	28.2%	15.3%
Rec_Afternoon: 10:00 a.m.–5:00 p.m.	2.4%	5.2%	26.9%	38.8%	26.7%
Rec_Evening: 6:00 p.m.–12:00 a.m.	5.5%	9.0%	26.6%	36.6%	22.3%

at different times of the day, prompting respondents to think about the impact of daily activities on warning comprehension.

We operationalize tornado warning response by asking participants to both retrospectively assess how often they have taken protective action in response to past tornado warnings and prospectively assess how confident they are that they will take protective action in response to future tornado warnings in a variety of circumstances. The items we use are shown in Table 5. The first item examines retrospective responsiveness to previous warnings among respondents who recall receiving at least one tornado warning (62% of the sample) and prospective responsiveness to future warnings among respondents who do not recall receiving a warning in the past (38% of the sample). Note that while the two groups are responding to different prompts, they address the same statement about protective action, so we treat them as equivalent in the analysis that follows. The next set of items measure prospective responsiveness in different situations, and the remaining items assess prospective responsiveness by time of day. These questions and the prompts that precede them are the same for respondents who do and do not recall receiving a warning in the past.

4. Reliability

a. Dimensionality

We use a combination of exploratory factor analysis and item response theory (IRT) to assess the dimensionality

TABLE 3. Operationalization of tornado warning comprehension (objective). Asterisks (*) indicate correct responses. Of the 55.3% of respondents who answered Ocomp_Warn_Time correct, 25.3% also correctly knew tornado warnings have an average lead time less than 30 min; of the 63.7% of respondents who answered Ocomp_Watch_Time correct, 27.7% also correctly knew tornado watches have an average lead time of 1 to 3 h.

Survey question	% correct
Ocomp_WW_Difference**:** 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. [*50% of participants get this version of the question.*]	
Tornado watch*	76.9%
Tornado warning	19.3%
Do not know	3.8%
Ocomp_Watch_Warn: This alert is used when a tornado is imminent. When this alert is issued, seek safe shelter immediately. [*50% of participants get this version of the question.*]	
Tornado watch	13.4%
Tornado warning*	84.0%
Do not know	2.6%
Ocomp_Warn_Time: If the National Weather Service issues a tornado warning for your area, how much time do you have before the tornado arrives?	
Less than 1 h*	55.3%
How many minutes are there between when tornado WARNINGS are issued and when tornadoes arrive? [<30 min]*	
1 to 24 h	36.6%
1 to 3 days	5.8%
More than 3 days	2.3%
Ocomp_Warn_Size: Approximately how large is the area included in an average tornado warning?	
Around the size of a city*	30.2%
Around the size of a county*	38.3%
Around the size of multiple counties	28.0%
Around the size of a state	2.2%
Around the size of multiple states	1.3%
Ocomp_Watch_Time: If the National Weather Service issues a tornado watch for your area, how much time do you have before the tornado arrives?	
Less than 1 h	19.1%
1 to 24 h*	63.7%
How many *hours* are there between when tornado watches are issued and when tornadoes arrive? [1 to 3 h]*	
1 to 3 days	13.1%
More than 3 days	4.1%
Ocomp_Watch_Size: Approximately how large is the area included in an average tornado watch?	
Around the size of a city	21.5%
Around the size of a county	33.3%
Around the size of multiple counties*	39.2%
Around the size of a state*	4.2%
Around the size of multiple states*	1.9%

and reliability of these measures. Factor analysis allows us to assess the dimensionality of each measure. We assume that the items we use to operationalize each concept (reception, objective comprehension, subjective comprehension, and response) coalesce to capture a single latent dimension for each concept. This is an important assumption because it allows us to combine the items into internally consistent scales that measure the underlying concepts. We test this assumption in two ways. First, we use scree plots to identify the number of dimensions (factors) that underlie the reception, subjective comprehension, and response measures.

As shown in Fig. 1, the scree plots clearly indicate a single factor solution for the measures. In all three cases, the line in the scree plots flattens to roughly match the *k*th factor (i.e., the 8th factor) at 2+ factors. This indicates that the first factor explains the plurality of variance in the items, whereas the remaining factors contribute little to the solution. Likewise, only the first factors have eigenvalues above 1.0, the limit that many researchers use to define acceptable factors (Kaiser 1960). This suggests that a single latent dimension underlies each of these measures.

In theory, five items may provide enough information to measure multiple dimensions of a concept; however, the magnitude and structure of the correlations necessary to do so are quite rare. Accordingly, we use correlations to evaluate inter-item consistency within the objective comprehension measure in place of a scree plot that explores the possibility of multiple dimensions. Table 6 shows the tetrachoric correlation matrix for the items we use to measure objective comprehension. It reveals a positive and statistically significant correlation between all but one of the items (Ocomp_Warn_Size). There is a significant negative correlation between comprehension of warning size and watch size (Ocomp_Watch_Size), no correlation between comprehension of warning size and knowledge of the difference between watches and warnings (Ocomp_WW_Difference), and positive but small correlations between warning size comprehension and the remaining items. This suggests that knowledge about the approximate geographic area covered by a tornado warning does *not* provide a reliable indication of tornado warning comprehension *along this dimension*. While comprehensive analysis is outside the scope of this study, this result may provide preliminary evidence that tornado warning comprehension includes multiple dimensions. Regardless of the reason, we do *not* include this item in the analysis that follows; when we remove it, the positive correlations that remain suggest that a single latent dimension underlies the

TABLE 4. Operationalization of tornado warning comprehension (subjective). For Scomp_WW_Difference: 1—Definitely no, 2—Probably no, 3—Not sure, 4—Probably yes, and 5—Definitely yes; for Scomp_WW_Understanding to Scomp_Radar: 1—Poor, 2—Fair, 3—Good, 4—Very good, and 5—Excellent; for Scomp_Morning to Scomp_Evenin: 1—Not at all confident, 2—Not very confident, 3—Somewhat confident, 4—Very confident, and 5—Extremely confident.

Survey question	1	2	3	4	5
Scomp_WW_Difference: Now we have some questions about the National Weather Service (NWS), an agency of the U.S. government that issues weather forecasts and different kinds of alerts to the public about hazardous weather, including severe weather watches and warnings. In general, do you understand the difference between watches and warnings?	1.3%	4.0%	10.2%	42.6%	41.9%
Scomp_WW_Understanding: How would you rate your understanding of tornado watches and warnings?	9.8%	27.5%	31.9%	21.6%	9.3%
Scomp_Severe_Thund: How would you rate your understanding of severe thunderstorm watches and warnings?	4.8%	23.7%	35.5%	25.0%	11.0%
Scomp_Maps: Forecasters, websites, and phone applications often use maps to display tornado watches and warnings. How would you rate your understanding of maps?	5.5%	17.7%	34.0%	26.4%	16.4%
Scomp_Radar: Forecasters, websites, and phone applications also use radar images to communicate tornado risk. How would you rate your understanding of radar images?	8.2%	23.6%	32.2%	24.3%	11.6%
For some people the time of day influences tornado warning reception, understanding, and/or responsiveness. If a tornado WARNING were issued for your area tomorrow at [RANDOM TIME], how confident are you that you would understand the warning?					
Scomp_Morning: 1:00 a.m.–9:00 a.m.	4.0%	9.7%	28.7%	38.5%	19.1%
Scomp_Afternoon: 10:00 a.m.–5:00 p.m.	2.3%	4.7%	25.2%	41.7%	26.1%
Scomp_Evening: 6:00 p.m.–12:00 a.m.	3.4%	6.4%	26.0%	39.9%	24.3%

remaining items.[2] This provides the remaining evidence we require to combine the items in each set into single scales that measure reception, subjective comprehension, objective comprehension, and response.

b. Discrimination

IRT allows us to assess the properties of each scale. How much information does each item contribute to the respective scales? And, more importantly, do the scales reliably capture the full range of participant ''ability'' in the sample? Said differently, do they adequately discriminate between people with low, average, and high reception, comprehension, and response tendencies? We address these questions by fitting IRT models to each of the scales. We fit graded response models (GRMs) to the polytomous reception, subjective comprehension, and response scales, and a two-parameter logistic model (2PLM) to the binary objective comprehension scale [see Edwards (2009) for an introduction to these models]. Rather than discussing the estimates for each item in the different scales, we focus on the scales as a whole and the information they convey across the range of ability. We do this by plotting the test information functions for each of the scales. If the tests are sufficiently discriminant, the distribution of information in each scale will be roughly symmetric across the range of ability, with a peak at zero (average ability) and asymptotic tails that approach zero as ability departs from average.

[2] In addition to assessing inter-item correlations, we use modified parallel analysis for dichotomous response items (Drasgow and Lissak 1983) to assess the dimensionality of this scale. Consistent with the correlations, this analysis indicates multidimensionality when we include Ocomp_Warn_Size and unidimensionality when we remove it.

TABLE 5. Operationalization of tornado warning response. For the Resp_Always: 1—Strongly disagree, 2—Disagree, 3—Neither disagree nor agree, 4—Agree, and 5—Strongly agree; for Resp_Sleep to Resp_Evening: 1—Not at all confident, 2—Not very confident, 3—Somewhat confident, 4—Very confident, and 5—Extremely confident.

Survey question	1	2	3	4	5
Please tell us how strongly you agree with the following statements about tornado WARNINGS. If you have never received a tornado WARNING, please tell us how you think you will respond if you receive a WARNING in the future:					
Resp_Always: I always take protective action when tornado warnings are issued for my area.	3.8%	14.0%	34.2%	34.9%	13.1%
Sometimes people receive tornado WARNINGS but *do not take protective action* because they are busy or doing something that makes it difficult to respond. For example, people often decide not to take protective action in response to tornado warnings when they are sleeping. How confident are you that you would *take protective action in response* to tornado warnings in the following situations?					
Resp_Sleep: If you are sleeping?	18.1%	29.9%	25.9%	16.1%	9.9%
Resp_Car: If you are in a car?	5.4%	17.2%	37.6%	25.8%	14.1%
Resp_Work: If you are at work or school?	3.3%	8.8%	32.0%	35.7%	20.2%
Resp_Store: If you are at a store?	4.1%	15.5%	38.7%	27.4%	14.2%
Resp_Sm_Group: If you are with a small group of friends or family?	2.9%	10.9%	36.8%	33.3%	16.1%
Resp_Lg_Group: If you are with a large group of friends or family?	3.4%	12.2%	34.9%	33.2%	16.3%
For some people the time of day influences tornado warning reception, understanding, and/or responsiveness. If a tornado WARNING were issued for your area tomorrow at [RANDOM TIME], how confident are you that you would take protective action in response to the warning?					
Resp_Morning: 1:00 a.m.–9:00 a.m.	6.0%	13.3%	35.7%	29.5%	15.5%
Resp_Afternoon: 10:00 a.m.–5:00 p.m.	2.9%	6.9%	33.2%	36.1%	20.8%
Resp_Evening: 6:00 p.m.–12:00 a.m.	4.7%	10.0%	33.2%	32.5%	19.5%

Figure 2a plots full test information functions for each of the scales we describe above. For reference, Fig. 2b plots the information function from the first item in each scale alone.[3] As Fig. 2a indicates, the full test information functions are generally symmetric, demonstrating high discrimination across the range of ability. The same is true of the item information functions for *Rec_All* and *Resp_Always*. By contrast, the first item information functions for *Ocomp_WW_Difference* and *Scomp_WW_Difference* are relatively asymmetric with a negative bias, indicating that these items alone do a poor job discriminating among high-ability participants. The negative bias in *Ocomp_WW_Difference*, for example, shows that correctly noting the difference between tornado watches and warnings is relatively easy for most participants. As such, it helps researchers differentiate between people of very low and low ability, but provides little information that differentiates between people with average and above average levels of comprehension. Fortunately, the symmetric shape of the full test

[3] Table A1 in the appendix provides more information on the discrimination of each item and scale by showing the percentage of information in each item/scale on the left and right side of each function.

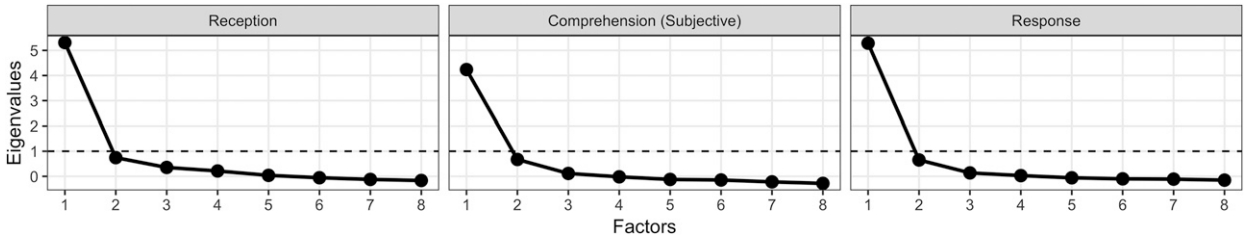

FIG. 1. Scree plots showing the eigenvalue of each latent dimension (factor number) in the tornado warning reception, subjective comprehension, and response item sets.

information function indicates that the other items in the objective comprehension scale make up for this lack of discrimination. The same can be said for the *Scomp_WW_Difference* and the full subjective comprehension scale—the first item insufficiently discriminates across the range of ability, but the remaining items in the scale make up for it.

These findings demonstrate the importance of multi-item scales and, simultaneously, enhance our confidence in the reliability of our measures. Accordingly, we use the IRT models to compute scale z scores (scale scores) for each participant in the sample. Like all z scores, 0 indicates average ability, and positive (negative) departures from 0 indicate standard deviations above (below) average. For instance, a respondent who scores a 1.2 on the objective comprehension scale is 1.2 standard deviations above average. These scores and the distributions they indicate provide consistent, transparent, replicable, and reliable measures of tornado warning reception, comprehension, and response that permit generalizable inference to populations of interest, be that the entire U.S. population, a specific region of the country, or a specific demographic group. The next section highlights this feature by exploring the sensitivity of the measures to geographic and demographic variation to identify significant differences that require attention in measurement. Future work will do the same by exploring variation in the measures over time.

5. Geographic and demographic variation in reception, comprehension, and response

a. Tornado warning exposure (geography)

We explore variation in reception, comprehension, and response by comparing them across geographic and demographic groups. In some cases, previous research and theory provide strong guidance on what we should observe. For example, we expect that tornado warning reception, comprehension, and response will vary as a function of exposure to tornado warnings. We examine this expectation by comparing scale scores to average yearly tornado warning counts in respondent's County Warning Areas (CWAs). We use six years of data from the NWS Performance Branch (1 January 2013–31 December 2018) to calculate these averages. Average warning counts range from 0 to 78 warnings per year during this time span and exhibit a slight positive skew because a relatively small set of CWAs issue significantly more warnings than other CWAs. If previous research and theory are correct, we will observe a positive relationship between respondent scale scores and tornado warning exposure. People who live in areas that frequently get tornado warnings will develop habits, coping mechanisms, and strategies that may make them more likely to receive, understand, and, possibly, take protective action in response to tornado warnings than people who rarely experience them. However, the overall relationship may be complicated because routine exposure to tornado warnings may encourage normalcy bias and complacency in some populations, whereas rarity of exposure might incite alarm and action (e.g., Simmons and Sutter 2009; Schultz et al. 2010; Ripberger et al. 2015b; Trainor et al. 2015). It is also possible that different geographic subcultures [e.g., Sims and Baumann 1972; however, see Davies-Jones et al. (1973) for a criticism] or unique geographic clusters of vulnerability that happen to overlap with warning frequency (e.g., Ashley 2007) will influence warning reception, comprehension, and response.

b. Demographics

Previous work provides less consistent guidance on the relationship between demographics and tornado warning reception, comprehension, and response. Some studies observe differences by gender, age, ethnicity, and race, whereas others do not. This inconsistency likely stems from multiple sources. It is possible that different studies produce different results because "true" differences between demographic groups are essentially zero and therefore difficult to distinguish from sampling variation. For example, some suggest that this is the case in hurricane evacuation research (e.g., Huang et al. 2016). Alternatively, there may be relatively

TABLE 6. Tetrachoric correlation matrix of objective tornado warning comprehension items. Standard errors are in parentheses; asterisks (*) indicate $p < 0.05$.

	Ocomp_ WW_Difference	Ocomp_ Warn_Time	Ocomp_ Warn_Size	Ocomp_ Watch_Time	Ocomp_ Watch_Size
Ocomp_WW_Difference	—	0.28 (0.04)*	0.00 (0.03)	0.17 (0.04)*	0.22 (0.03)*
Ocomp_Warn_Time	0.28 (0.04)*	—	0.13 (0.03)*	0.31 (0.03)*	0.19 (0.03)*
Ocomp_Warn_Size	0.00 (0.03)	0.13 (0.03)*	—	0.07 (0.03)*	−0.77 (0.02)*
Ocomp_Watch_Time	0.17 (0.04)*	0.31 (0.03)*	0.07 (0.03)*	—	0.06 (0.03)*
Ocomp_Watch_Size	0.22 (0.03)*	0.19 (0.03)*	−0.77 (0.02)*	0.06 (0.03)*	—

small but ''true'' differences that are observationally inconsistent due to variation in research design, data collection, and measurement in previous studies. We explore this possibility by comparing scale scores across demographic groups.

c. Methods

We use multiple linear regression to identify the comparisons we describe above. The models regress reception, objective comprehension, subjective comprehension, and response on CWA tornado warning count, gender, age, ethnicity (Hispanic), and race. To identify possible inflections and nonlinearities in the relationship between continuous explanatory variables (tornado warning count and age) and respondent scale scores, we include cubic polynomials (e.g., age + age^2 + age^3) for these terms. Table 7 lists the parameter estimates we obtain from these models. Figure 3 uses these estimates to plot the scores for each group on each scale when the covariates in the models are set to their respective means (for continuous covariates) and modes (for discrete covariates).[4]

d. Results

Figure 3a plots the effect of warning exposure on reception, objective comprehension, subjective comprehension, and response. Consistent with extant research and theory, there is a positive and statistically significant relationship between exposure and the scales. Note, however, that the shape of the relationship varies across the scales. The relationship for the first three scales indicates a ceiling effect, suggesting that there is relatively little difference in between people who live in moderate and high exposure areas. The magnitude of the relationship varies across the scales as well. Exposure has a relatively small effect on response. People who live in relatively low-frequency warning areas (5 tornado warnings per year) score close to average (0σ) on the response scale, whereas people who live in high-frequency areas (40 tornado warnings per year) score slightly above average (0.14σ), resulting in a difference of only 0.14σ between the two groups. Subjective comprehension is more sensitive to exposure; the difference between people who live in relatively low- and high-frequency warning areas is 0.35σ. Reception and objective comprehension are fairly sensitive as well; on both measures, the scale score differences are roughly 0.20σ between people who live in low- and high-frequency warning areas. In addition to providing important information about the relationship between exposure and warning reception, comprehension, and response, these findings are broadly consistent with previous research and theory.

Figure 3b plots the effect of age on tornado warning reception, objective comprehension, subjective comprehension, and response. As with exposure, the shape of the relationship between age and scale score varies by measure. The relationship is generally negative for reception, subjective comprehension, and response, but scores on these scales do not monotonically decrease with age; rather, they increase then decrease, suggesting relative deficiencies among young adults and the elderly and proficiencies among people in the middle-aged group. In terms of magnitude, age has the greatest effect on reception; the mean difference between middle-aged adults (35 years old) and older adults (60 years old) is 0.50σ, which is roughly twice the size of the difference between these two groups on the subjective comprehension and response scales. Contrasting these trends, age has the opposite effect on objective comprehension; it increases with age, but the difference between age groups is not as large (0.15σ). Nevertheless, it is worth noting that older groups have relatively high levels of objective comprehension, but little confidence that they will get, understand, and take protective action in response to tornado warnings.

[4] In this case, the ''average'' person is a non-Hispanic White female, who is 46 years old, and lives in a CWA that issues approximately 17 tornado warnings per year. The panels in Fig. 3 show the change in ability that occurs when we vary each of these parameters in isolation while holding the others constant at their average values.

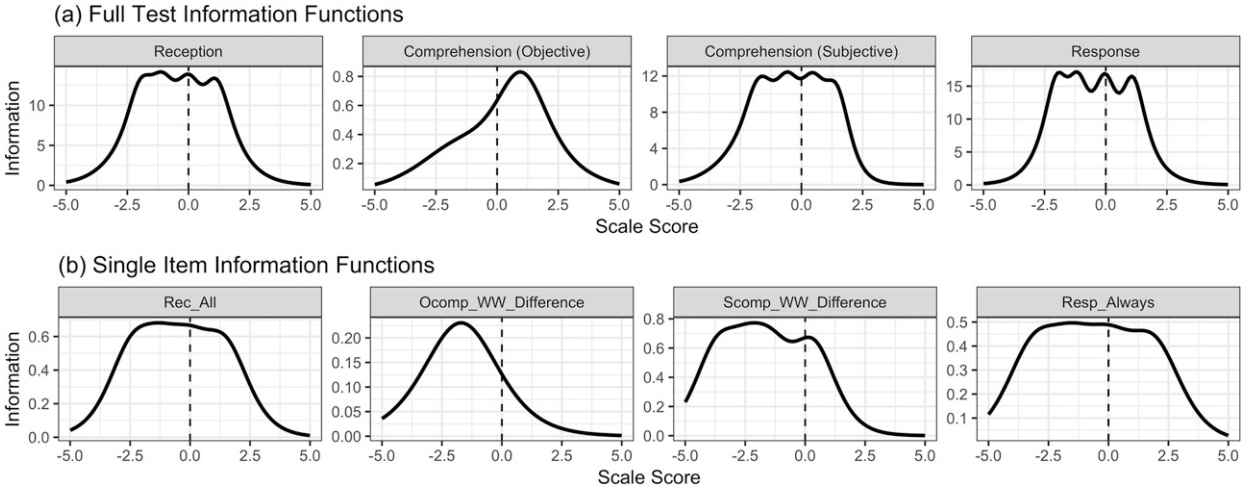

FIG. 2. Example single item and full test information functions from the item response theory (IRT) models of tornado warning reception, objective comprehension, subjective comprehension, and response.

Figure 3c plots the effect of gender, ethnicity, and race on tornado warning reception, objective comprehension, subjective comprehension, and response. Consistent with the coefficient estimates in Table 7, there is no difference between men and women in reception, objective comprehension, and response. There is, however, a significant difference (0.26σ) in subjective comprehension; men have more confidence in warning comprehension than women, despite comparable levels of objective comprehension. Similarly, there is little difference between non-Hispanic and Hispanic respondents on the reception, subjective comprehension, and response scales, but there is a small difference (0.14σ) on the objective comprehension scale, where Hispanic respondents score (on average) a bit lower on the scale than non-Hispanic groups. Finally, there are multiple interesting differences across the race groups. In comparison to people who identify as White, respondents who identify as Black or African American score higher on the reception and response scales, but lower on the objective and subjective comprehension scales. Those who identify with a different race (e.g., Asian) or two or more races score significantly lower than White (and Black) groups on all four scales. The difference is most noticeable on the subjective comprehension scale, where the difference between people who identify as White and people who identify with ''Other'' race groups is 0.44σ.

Overall, these findings suggest significant differences in tornado warning reception, comprehension,

TABLE 7. Parameter estimates from linear regression models. Linear regression coefficients; standard errors in parentheses; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

	Reception	Objective comprehension	Subjective comprehension	Response
Male (vs. female)	0.04 (0.04)	0.02 (0.02)	0.26*** (0.03)	−0.004 (0.04)
Age[1]	−9.52*** (0.96)	4.06*** (0.58)	−2.31** (0.91)	−4.36*** (0.97)
Age[2]	−2.87*** (0.95)	−2.76*** (0.57)	−3.82*** (0.90)	−2.57*** (0.96)
Age[3]	3.10*** (0.95)	−2.33*** (0.57)	2.88*** (0.90)	2.69*** (0.96)
Hispanic (vs. non-Hispanic)	0.07 (0.05)	−0.08*** (0.03)	−0.05 (0.04)	0.08* (0.05)
Black (vs. White)	0.16*** (0.05)	−0.21*** (0.03)	−0.21*** (0.05)	0.14*** (0.05)
Other Race (vs. White)	−0.17*** (0.06)	−020.*** (0.04)	−0.44*** (0.06)	−0.22*** (0.06)
Warning count[1]	4.76*** (0.95)	4.98*** (0.58)	9.00*** (0.90)	2.49** (0.97)
Warning count[2]	−3.21*** (0.95)	−2.12*** (0.57)	−4.91*** (0.90)	−0.89 (0.96)
Warning count[3]	0.70 (0.95)	1.54** (0.57)	2.54** (0.90)	−0.84 (0.96)
Constant	−0.05* (0.03)	0.08*** (0.02)	−0.04 (0.03)	−0.02 (0.03)
Observations	3000	3000	3000	3000
R^2	0.06	0.08	0.10	0.02
Adjusted R^2	0.05	0.08	0.10	0.02
Residual standard error	0.95	0.57	0.90	0.96

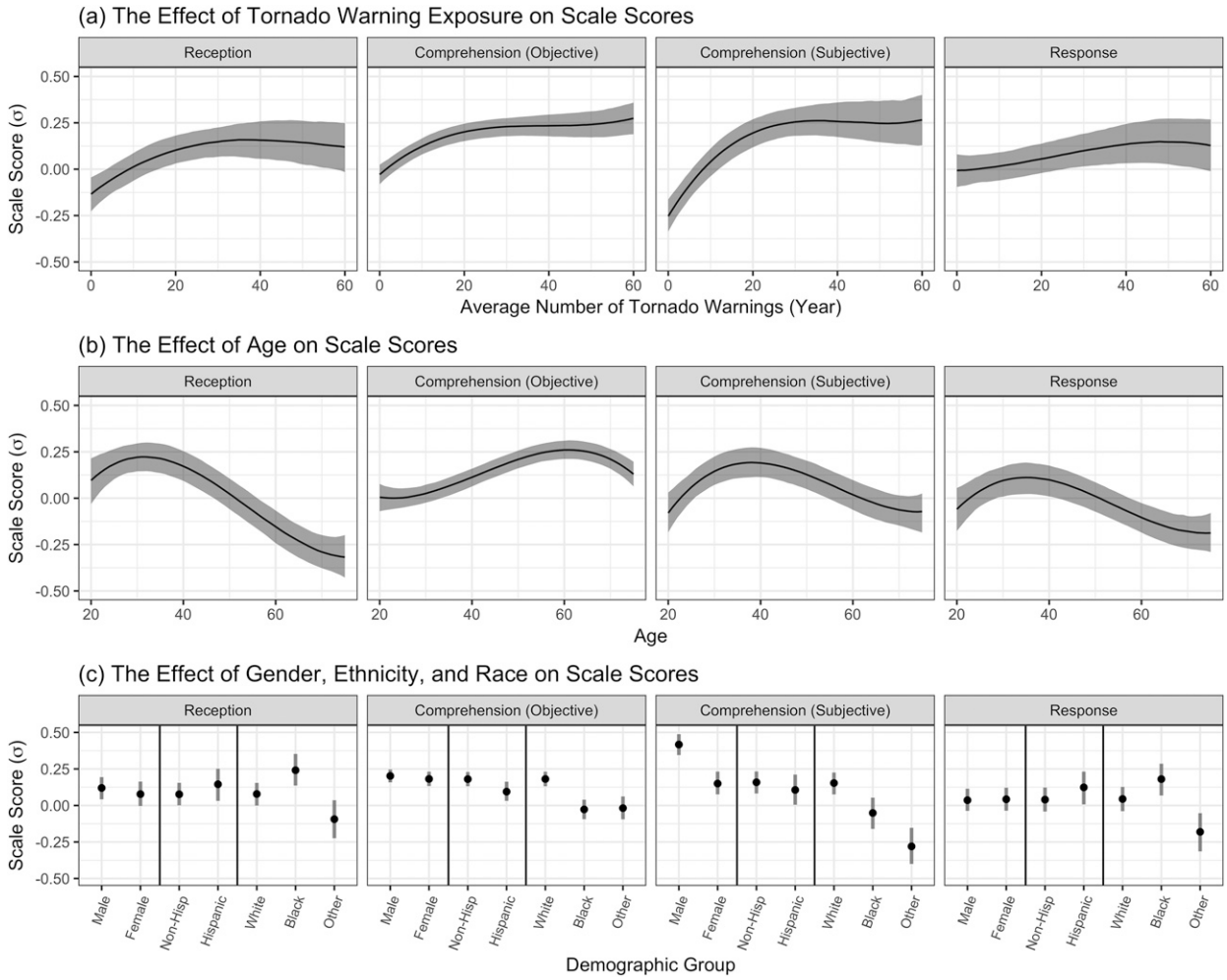

FIG. 3. Predictions from the linear regression models as one input varies and the others remain constant at average values.

and response across the United States. In addition to advancing research on the geographic and demographic correlates of these concepts, these findings strongly indicate that efforts to include social criteria in performance management and evaluation must be sensitive to the geographic and demographic characteristics of the population. That is, efforts to characterize the whole population with a single estimate are likely to miss important sources of variation across communities.

6. Discussion

The mission of the NWS is to provide weather, water, and climate data, forecasts, and warnings for the protection of life and property and the enhancement of the national economy. In addition to forecast and warning verification statistics, we follow recent reports by NOAA and the National Academies of Sciences (NAS) (NOAA 2015; National Academies of Sciences, Engineering, and Medicine 2018) in advocating the development and use of social criteria in performance management and evaluation. Doing so will require dialogue between policy makers and practitioners—the people who will employ the social criteria—and the researchers who produce them. Important topics of conversation will need to include relevant criteria (e.g., what concepts require measurement?) and appropriate methodologies that produce consistent, transparent, replicable, and reliable measures of these criteria that are generalizable to populations of interest. We hope that this study will advance this dialogue by suggesting three criteria that require measurement (forecast and warning reception, comprehension, and response) and introducing a methodology that allows us to systematically measure these concepts over time and space in a single information domain (tornado warnings). The methodology

we employ improves upon previous research on tornado warning reception, comprehension, and response in multiple ways. It provides a more a generalizable approach to measurement using a temporally consistent set of survey questions that are applicable across the United States and it is more transparent and replicable than previous research because the data and methods (source code) are publicly available.

The methodology we propose also improves upon previous research by systematically analyzing the reliability of the measures we develop. The psychometric tests we use to assess reliability suggest that the measures capture single dimensions and adequately discriminate between people with low, average, and high tornado warning reception, comprehension, and response tendencies. Demonstrating this type of reliability is imperative to developing performance management and evaluation measures. If the measures are not reliable, they will not provide reliable feedback on how to improve the system.

In addition to describing and assessing the reliability of the measures, we explore the sensitivity of the measures to geographic and demographic variation across the population. Consistent with previous research and theory, the measures show predictable differences in exposure across the country. People who live in geographic areas that rarely get tornado warnings demonstrate relatively low levels of reception and comprehension in comparison to people who live in high volume warning areas. This gives us confidence that the measures are tapping the constructs they purport to be measuring. Perhaps more importantly, the regression models corroborate and in many cases clarify previous research on the influence of demographic differences on tornado warning reception, comprehension, and response. Most notably, the models suggest that geographic and demographic differences, especially in age and race, have significant effects on tornado warning reception, comprehension, and response that will require attention when tracking these measures over time and space for performance measurement and evaluation.

7. Next steps

We believe that this study represents an important first step toward the development and use of social criteria in performance management and evaluation, but recognize that the current study faces many limitations that suggest avenues for future research. One of the most obvious limitations of this study is the focus on a single type of information that the NWS produces, namely tornado warnings. Future research is necessary to develop reliable measures of reception, comprehension, and response across domains. This may require modification of the methodology as national surveys are not necessarily appropriate for all types of information (e.g., coastal/lakeshore hazards).

In addition to a narrow information domain, this study includes a somewhat narrow set of criteria for measurement (reception, comprehension, and response). We believe that these are among the most important but are not the *only* factors that impact the link between NWS information provision and public response. Future research and dialogue are therefore necessary to identify additional criteria that may require measurement. Candidate criteria might include public risk perceptions, trust in information from the NWS and/or public awareness of extreme weather, climate, and water impacts and how to prevent or mitigate those impacts (Allan et al. 2017). More work on forecast and warning reception, comprehension, and response may be necessary as well. For example, future studies might work to develop a more comprehensive set of items that measure multiple dimensions of tornado warning comprehension, such as public judgements of the probability that a tornado will strike given a warning polygon (Jon et al. 2018, 2019).

The methodology we use in the study is also subject to multiple limitations. Most notably, it is possible that the estimates we get from large Internet surveys are not generalizable to segments of the U.S. population who do not use the Internet. Recent estimates suggest that roughly 11% of U.S. adults *do not* use the Internet. This group disproportionately includes people above the age of 65, people who make less than $30,000 a year, people with less than a high school education, and people who live in rural areas (Pew Research Center 2018). We therefore advocate the use of multiple quantitative and qualitative methodologies, including non-internet surveys, interviews, and focus groups, that target these populations when measuring concepts like forecast and warning reception, comprehension, and response.

It is also important to note that the methodology we use in this study provides baseline estimates of tornado warning, reception, comprehension, and response across the country. Using this information to evaluate and improve NWS operations over time and across jurisdictions will require 1) time series data and 2) small area estimates. Time series data will allow us to measure the impact of significant policy changes on warning reception, comprehension, and response. This is particularly urgent as the NWS considers large-scale changes to the information environment, such as Hazard Simplification (NWS 2018) and the Forecasting a Continuum of

TABLE A1. Item and test information functions and distributions.

Item	Item information	Information on the negative side of the function	Information on the positive side of the function
Rec_All: I receive all tornado warnings that are issued for my area.	4.04	2.36 (58.38%)	1.68 (41.62%)
Rec_Most: I receive most tornado warnings that are issued for my area.	2.86	1.77 (61.72%)	1.1 (38.28%)
Rec_Soon: I receive tornado warnings as soon as they are issued for my area.	4.32	2.59 (59.97%)	1.73 (40.03%)
Rec_Sleep: If you are sleeping?	3.87	1.5 (38.75%)	2.37 (61.25%)
Rec_Car: If you are in a car?	6.10	3.44 (56.42%)	2.66 (43.58%)
Rec_Work: If you are at work or school?	6.36	3.98 (62.6%)	2.38 (37.4%)
Rec_Store: If you are at a store?	7.91	4.35 (55.02%)	3.56 (44.98%)
Rec_Sm_Group: If you are with a small group of friends or family?	11.79	7.13 (60.53%)	4.65 (39.47%)
Rec_Lg_Group: If you are with a large group of friends or family?	9.28	5.69 (61.25%)	3.6 (38.75%)
Rec_Morning: 1:00 a.m.–9:00 a.m.	3.67	1.95 (53.06%)	1.72 (46.94%)
Rec_Afternoon: 10:00 a.m.–5:00 p.m.	4.74	3.22 (67.94%)	1.52 (32.06%)
Rec_Evening: 6:00 p.m.–12:00 a.m.	3.78	2.38 (62.9%)	1.4 (37.1%)
TORNADO WARNING RECEPTION TEST (TOTAL), alpha = 0.90	68.72	40.35 (58.72%)	28.36 (41.28%)
Ocomp_WW_Difference: Watch vs warning	1.23	1.05 (83.66%)	0.18 (16.34%)
Ocomp_Warn_Time: Warning time	1.07	0.22 (18.00%)	0.85 (82.00%)
Ocomp_Warn_Size: Warning size	—	—	—
Ocomp_Watch_Time: Watch time	0.77	0.34 (25.99%)	0.43 (74.01%)
Ocomp_Watch_Size: Watch size	0.61	0.28 (45.44%)	0.33 (54.56%)
OBJECTIVE TORNADO COMPREHENSION TEST (TOTAL), alpha = 0.51[a]	3.69	1.9 (40.23%)	1.8 (59.77%)
Scomp_WW_Difference: [. . .] do you understand the difference between watches and warnings?	4.44	3.42 (77.12%)	1.02 (22.88%)
Scomp_WW_Understanding: [. . .] understanding of tornado watches and warnings?	10.24	4.97 (48.5%)	5.27 (51.5%)
Scomp_Severe_Thund: [. . .] understanding of severe thunderstorm watches and warnings?	11.35	6.09 (53.65%)	5.26 (46.35%)
Scomp_Maps: [. . .] understanding of maps?	7.53	4.27 (56.75%)	3.26 (43.25%)
Scomp_Radar: [. . .] understanding of radar images?	8.66	4.5 (52.03%)	4.15 (47.97%)
Scomp_Morning: 1:00 a.m.–9:00 a.m.	5.43	3.36 (61.89%)	2.07 (38.11%)
Scomp_Afternoon: 10:00 a.m.–5:00 p.m.	5.70	3.87 (67.84%)	1.83 (32.16%)
Scomp_Evening: 6:00 p.m.–12:00 a.m.	5.27	3.48 (66.06%)	1.79 (33.94%)
SUBJECTIVE TORNADO COMPREHENSION TEST (TOTAL), alpha = 0.90	58.63	33.97 (57.95%)	24.65 (42.05%)
Resp_Always: I always take protective action when tornado warnings are issued for my area.	3.68	2.15 (58.5%)	1.53 (41.5%)
Resp_Sleep: If you are sleeping?	4.05	1.81 (44.78%)	2.24 (55.22%)
Resp_Car: If you are in a car?	6.48	3.66 (56.41%)	2.83 (43.59%)
Resp_Work: If you are at work or school?	7.86	5.0 (63.55%)	2.87 (36.45%)
Resp_Store: If you are at a store?	9.69	5.6 (57.77%)	4.09 (42.23%)
Resp_Sm_Group: If you are with a small group of friends or family?	14.56	9.08 (62.41%)	5.47 (37.59%)

TABLE A1. (*Continued*)

Item	Item information	Information on the negative side of the function	Information on the positive side of the function
Resp_Lg_Group: If you are with a large group of friends or family?	12.46	7.71 (61.9%)	4.75 (38.1%)
Resp_Morning: 1:00 a.m.–9:00 a.m.	5.37	3.07 (57.24%)	2.3 (42.76%)
Resp_Afternoon: 10:00 a.m.–5:00 p.m.	6.52	4.2 (64.5%)	2.31 (35.5%)
Resp_Evening: 6:00 p.m.–12:00 a.m.	5.66	3.53 (62.35%)	2.13 (37.65%)
TORNADO WARNING RESPONSE TEST (TOTAL), alpha = 0.91	76.32	45.82 (60.03%)	30.51 (39.97%)

[a] Calculated using the tetrachoric correlation matrix in Table 6.

Environmental Threats (FACETs) initiative (Rothfusz et al. 2018). Absent consistent and statistically reliable data before and after changes of this magnitude, it is extremely difficult to evaluate the effectiveness of these changes. Small area estimates will allow us to compare reception, comprehension, and response across NWS regions and forecast offices with hopes of empirically identifying best practices in community engagement (Ripberger et al. 2019). This feedback will be extremely valuable as NWS regions and offices initiate and assess education campaigns and develop training materials that attempt to improve risk communication and community response to weather, climate, and water information.

Finally, we note that information of this sort will *only* improve NWS performance if practitioners (e.g., forecasters) and managers use it when making decisions. In transitioning this information from research to NWS operations, it is imperative that we, as a community, develop tools, routines, and procedures that facilitate and institutionalize information utilization. While there are many ways to accomplish this, one approach may involve the incorporation of this information in a training course such as those offered by the NWS Warning Decision Training Division. A complementary approach may involve the curation and dissemination of this information by the NWS Performance Management Branch. Regardless of the specific approach, we believe the incorporation of this information in NWS operations will help the organization better achieve its mission to protect life and property and enhance the national economy.

APPENDIX

Item Response Theory (IRT) Summary Statistics

Table A1 provides more on the IRT models we estimate by showing the information that each item contributes to the scales and the percentage of information in each item/scale on the left and right side of each function.

REFERENCES

Aguirre, B., 1988: The lack of warnings before the Saragosa tornado. *Int. J. Mass Emerg. Disasters*, **6**, 65–74.

Allan, J. N., J. T. Ripberger, V. T. Ybarra, and E. Cokely, 2017: The Oklahoma Warning Awareness Scale: A psychometric analysis of a brief self-report survey instrument. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, **61**, 1203–1207, https://doi.org/10.1177/1541931213601783.

Ashley, W. S., 2007: Spatial and temporal analysis of tornado fatalities in the United States: 1880–2005. *Wea. Forecasting*, **22**, 1214–1228, https://doi.org/10.1175/2007WAF2007004.1.

Balluz, L., L. Schieve, T. Holmes, S. Kiezak, and J. Malilay, 2000: Predictors for people's response to a tornado warning: Arkansas, 1 March 1997. *Disasters*, **24**, 71–77, https://doi.org/10.1111/1467-7717.00132.

Biddle, M. D., 2007: Warning reception, response, and risk behavior in the 3 May 1999 Oklahoma City long-track violent tornado. Ph.D. thesis, University of Oklahoma, 140 pp., https://shareok.org/handle/11244/1289.

Brown, S., P. Archer, E. Kruger, and S. Mallonee, 2002: Tornado-related deaths and injuries in Oklahoma due to the 3 May 1999 tornadoes. *Wea. Forecasting*, **17**, 343–353, https://doi.org/10.1175/1520-0434(2002)017<0343:TRDAII>2.0.CO;2.

Casteel, M. A., 2016: Communicating increased risk: An empirical investigation of the National Weather Service's impact-based warnings. *Wea. Climate Soc.*, **8**, 219–232, https://doi.org/10.1175/WCAS-D-15-0044.1.

——, 2018: An empirical assessment of impact based tornado warnings on shelter in place decisions. *Int. J. Disaster Risk Reduct.*, **30**, 25–33, https://doi.org/10.1016/j.ijdrr.2018.01.036.

Chaney, P. L., and G. S. Weaver, 2010: The vulnerability of mobile home residents in tornado disasters: The 2008 Super Tuesday tornado in Macon County, Tennessee. *Wea. Climate Soc.*, **2**, 190–199, https://doi.org/10.1175/2010WCAS1042.1.

——, ——, S. A. Youngblood, and K. Pitts, 2013: Household preparedness for tornado hazards: The 2011 disaster in DeKalb County, Alabama. *Wea. Climate Soc.*, **5**, 345–358, https://doi.org/10.1175/WCAS-D-12-00046.1.

Cohen, R. J., M. E. Swerdlik, and S. M. Phillips, 1996: *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. 3rd ed. Mayfield Publishing Co., 752 pp.

Comstock, R. D., and S. Mallonee, 2005: Comparing reactions to two severe tornadoes in one Oklahoma community. *Disasters*, **29**, 277–287, https://doi.org/10.1111/j.0361-3666.2005.00291.x.

Davies-Jones, R., J. Golden, J. Schaefer, R. H. Pine, H. E. Landsberg, L. Pedersen, J. H. Sims, and D. D. Baumann, 1973: Psychological response to tornadoes. *Science*, **180**, 544–548, https://doi.org/10.1126/science.180.4086.544.

Drabek, T. E., 1986: *Human System Responses to Disaster: An Inventory of Sociological Findings*. Springer, 509 pp.

Drasgow, F., and R. I. Lissak, 1983: Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *J. Appl. Psychol.*, **68**, 363, https://doi.org/10.1037/0021-9010.68.3.363.

Edwards, M. C., 2009: An introduction to item response theory using the Need for Cognition Scale. *Soc. Personal. Psychol. Compass*, **3**, 507–529, https://doi.org/10.1111/j.1751-9004.2009.00194.x.

Godfrey, C. M., P. Wolf, M. K. Goldsbury, J. A. Caudill, and D. P. Wedig, 2011: An evaluation of convective warning utilization by the general public. *39th Conf. on Broadcast Meteorology,* Oklahoma City, OK, Amer. Meteor. Soc., 1.3, https://ams.confex.com/ams/39BROADCAST/webprogram/Paper189100.html.

Hammer, B., and T. W. Schmidlin, 2002: Response to warnings during the 3 May 1999 Oklahoma City tornado: Reasons and relative injury rates. *Wea. Forecasting*, **17**, 577–581, https://doi.org/10.1175/1520-0434(2002)017<0577:RTWDTM>2.0.CO;2.

Huang, S., M. Lindell, and C. S. Prater, 2016: Who leaves and who stays? A review and statistical meta-analysis of hurricane evacuation studies. *Environ. Behav.*, **48**, 991–1029, https://doi.org/10.1177/0013916515578485.

Jauernic, S. T., and M. S. Van Den Broeke, 2016: 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.

Jon, I., S.-K. Huang, and M. K. Lindell, 2018: Perceptions and reactions to tornado warning polygons: Would a gradient polygon be useful?. *Int. J. Disaster Risk Reduct.*, **30**, 132–144, https://doi.org/10.1016/j.ijdrr.2018.01.035.

——, ——, and ——, 2019: Perceptions and expected immediate reactions to severe storm displays. *Risk Anal.*, **39**, 274–290, https://doi.org/10.1111/risa.12896.

Kaiser, H. F., 1960: The application of electronic computers to factor analysis. *Educ. Psychol. Meas.*, **20**, 141–151, https://doi.org/10.1177/001316446002000116.

Klockow, K. E., 2013: Spatializing tornado warning lead-time: Risk perception and response in a spatio-temporal framework. Ph.D. thesis, University of Oklahoma, 284 pp.

Lindell, M. K., and R. W. Perry, 2012: The protective action decision model: Theoretical modifications and additional evidence. *Risk Anal.*, **32**, 616–632, https://doi.org/10.1111/j.1539-6924.2011.01647.x.

——, S. K. Huang, H. L. Wei, and C. D. Samuelson, 2016: Perceptions and expected immediate reactions to tornado warning polygons. *Nat. Hazards*, **80**, 683–707, https://doi.org/10.1007/s11069-015-1990-5.

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.

Mileti, D. S., and J. H. Sorensen, 1990: Communication of emergency public warnings: A social science perspective and state-of-the-art assessment. Oak Ridge National Laboratory Rep. ORNL-6609, 166 pp., http://www.cires.org.mx/docs_info/CIRES_003.pdf.

Miran, S. M., C. Ling, and L. Rothfusz, 2018: Factors influencing people's decision-making during three consecutive tornado events. *Int. J. Disaster Risk Reduct.*, **28**, 150–157, https://doi.org/10.1016/j.ijdrr.2018.02.034.

Mitchem, J. D., 2003: An analysis of the September 20, 2002 Indianapolis tornado: Public response to a tornado warning and damage assessment difficulties. Natl. Hazards Quick Response Rep. 161, 55 pp.

National Academies of Sciences, Engineering, and Medicine, 2018: *Integrating Social and Behavioral Sciences within the Weather Enterprise*. The National Academies Press, 198 pp., https://doi.org/10.17226/24865.

NOAA, 2015: Vision and strategy: Supporting NOAA's mission with social science. NOAA Tech. Rep., 19 pp., https://www.performance.noaa.gov/wp-content/uploads/SSVS_Final_073115.pdf.

NWS, 2018: Understand tornado alerts. National Weather Service, accessed 18 December 2018, https://www.weather.gov/safety/tornado-ww.

Paul, B. K., and M. Stimers, 2012: Exploring probable reasons for record fatalities: the case of 2011 Joplin, Missouri, Tornado. *Nat. Hazards*, **64**, 1511–1526, https://doi.org/10.1007/s11069-012-0313-3.

——, ——, and M. Caldas, 2015: Predictors of compliance with tornado warnings issued in Joplin, Missouri, in 2011. *Disasters*, **39**, 108–124, https://doi.org/10.1111/disa.12087.

Pew Research Center, 2018: 11% of Americans don't use the internet. Who are they? Pew Research Center, accessed 20 November 2018, https://www.pewresearch.org/fact-tank/2018/03/05/some-americans-dont-use-the-internet-who-are-they/.

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., https://ams.confex.com/ams/88Annual/webprogram/Paper132939.html.

Ripberger, J. T., C. L. Silva, H. C. Jenkins-Smith, and M. James, 2015a: The influence of consequence-based messages on public responses to tornado warnings. *Bull. Amer. Meteor. Soc.*, **96**, 577–590, https://doi.org/10.1175/BAMS-D-13-00213.1.

——, ——, ——, D. E. Carlson, M. James, and K. G. Herron, 2015b: False alarms and missed events: The impact and origins of perceived inaccuracy in tornado warning systems. *Risk Anal.*, **35**, 44–56, https://doi.org/10.1111/risa.12262.

——, J. Allan, W. W. Wehde, M. Krocak, C. L. Silva, and H. C. Jenkins-Smith, 2019: Tornado warning reception, comprehension, and response across county warning areas in the United States. *14th Symp. on Societal Applications: Policy Research and Practice*, Phoenix, AZ, Amer. Meteor. Soc., 3.5, https://ams.confex.com/ams/2019Annual/meetingapp.cgi/Paper/351769.

Robinson, S. E., J. M. Pudlo, and W. Wehde, 2019: The new ecology of tornado warning information: A natural experiment assessing threat intensity and citizen-to-citizen information sharing. *Public Admin. Rev.*, https://doi.org/10.1111/puar.13030, in press.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow, 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.

Schmidlin, T. W., B. O. Hammer, Y. Ono, and P. S. King, 2009: Tornado shelter-seeking behavior and tornado shelter options

among mobile home residents in the United States. *Nat. Hazards*, **48**, 191–201, https://doi.org/10.1007/s11069-008-9257-z.

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.

Senkbeil, J. C., M. S. Rockman, and J. B. Mason, 2012: Shelter seeking plans of Tuscaloosa residents for a future tornado event. *Wea. Climate Soc.*, **4**, 159–171, https://doi.org/10.1175/WCAS-D-11-00048.1.

Sherman-Morris, K., 2010: Tornado warning dissemination and response at a university campus. *Nat. Hazards*, **52**, 623–638, https://doi.org/10.1007/s11069-009-9405-0.

Silva, C. L., J. T. Ripberger, H. C. Jenkins-Smith, and M. Krocak, 2017: Establishing a baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. University of Oklahoma Center for Risk and Crisis Management, 30 pp., http://risk.ou.edu/downloads/news/WX17-Reference-Report.pdf.

——, ——, ——, ——, and W. W. Wehde, 2018: Refining the baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. University of Oklahoma Center for Risk and Crisis Management, 29 pp., http://risk.ou.edu/downloads/news/WX18-Reference-Report.pdf.

Silver, A., and J. Andrey, 2014: The influence of previous disaster experience and sociodemographics on protective behaviors during two successive tornado events. *Wea. Climate Soc.*, **6**, 91–103, https://doi.org/10.1175/WCAS-D-13-00026.1.

Simmons, K. M., and D. Sutter, 2009: False alarms, tornado warnings, and tornado casualties. *Wea. Climate Soc.*, **1**, 38–53, https://doi.org/10.1175/2009WCAS1005.1.

Sims, J. H., and D. D. Baumann, 1972: The tornado threat: Coping styles of the north and south. *Science*, **176**, 1386–1392, https://doi.org/10.1126/science.176.4042.1386.

Trainor, J. E., D. Nagele, B. Philips, and B. Scott, 2015: Tornadoes, social science, and the false alarm effect. *Wea. Climate Soc.*, **7**, 333–352, https://doi.org/10.1175/WCAS-D-14-00052.1.