LETTER • OPEN ACCESS

Environmental risk perception from visual cues: the psychophysics of tornado risk perception

To cite this article: Barry Dewitt et al 2015 Environ. Res. Lett. 10 124009

Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is © 2015 IOP Publishing Ltd.

As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 3.0 licence, this Accepted Manuscript is available for reuse under a CC BY 3.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence https://creativecommons.org/licenses/by/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the article online for updates and enhancements.
Manuscript version: Accepted Manuscript

The "Accepted Manuscript" is the author's original version of an article including any changes made following the peer review process but excluding any editing, typesetting or other changes made by IOP Publishing and/or its licensors.

During the embargo period (the 12 month period from publication of the Version of Record of this article), the Accepted Manuscript:

- is fully protected by copyright and can only be accessed by subscribers to the journal;
- cannot be reused or reposted elsewhere by anyone unless an exception to this policy has been agreed in writing with IOP Publishing.

As the Version of Record of this article is going to be/has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after a 12 month embargo period.

After the embargo period, everyone is permitted to copy and redistribute this article for Non-Commercial purposes only, provided they:

- give appropriate credit and provide the appropriate copyright notice;
- show that this article is published under a CC BY-NC-ND 3.0 licence;
- provide a link to the CC BY-NC-ND 3.0 licence;
- provide a link to the Version of Record;
- do not use this article for commercial advantage or monetary compensation; and
- only use this article in its entirety and do not make derivatives from it.

*Please see CC BY-NC-ND 3.0 licence for full terms.

View the Version of Record for this article online at iopscience.org.
Environmental risk perception from visual cues: Caution and sensitivity in evaluating tornado risks

Barry Dewitt
Department of Engineering & Public Policy, Carnegie Mellon University

Baruch Fischhoff
Department of Engineering & Public Policy and Department of Social & Decision Sciences, Carnegie Mellon University

Alexander Davis
Department of Engineering & Public Policy, Carnegie Mellon University

Stephen B. Broomell
Department of Social & Decision Sciences, Carnegie Mellon University
Abstract

Lay judgments of environmental risks are central to both immediate decisions (e.g., taking shelter from a storm) and long-term ones (e.g., building in locations subject to storm surges). Using methods from quantitative psychology, we provide a general approach to studying lay perceptions of environmental risks. As a first application of these methods, we investigate a setting where lay decisions have not taken full advantage of advances in natural science understanding: tornado forecasts in the US and Canada. Because official forecasts are imperfect, members of the public must often evaluate the risks on their own, by checking environmental cues (such as cloud formations) before deciding whether to take protective action. We study lay perceptions of cloud formations, demonstrating an approach that could be applied to other environmental judgments. We use signal detection theory to analyze how well people can distinguish tornadic from non-tornadic clouds, and multidimensional scaling to determine how people make these judgments. We find that participants (N = 400 recruited from Amazon Mechanical Turk) have heuristics that generally serve them well, helping participants to separate tornadic from non-tornadic clouds, but which also lead them to misjudge the tornado risk of certain cloud types. The signal detection task revealed confusion regarding shelf clouds, mammatus clouds, and clouds with upper- and mid-level tornadic features, which the multidimensional scaling task suggested was the result of participants focusing on the darkness of the weather scene and the ease of discerning its features. We recommend procedures for training (e.g., for storm spotters) and communications (e.g., tornado warnings) that will reduce systematic misclassifications of tornadicity arising from observers’ reliance on otherwise useful heuristics.

Keywords: decision-making, risk perception, risk communication, signal detection theory, multidimensional scaling
Environmental risk perception from visual cues: Caution and sensitivity in evaluating tornado risks

“If you hear a roaring sound, or see a funnel cloud, you should seek shelter immediately.”

— Canadian Broadcasting Corporation (CBC) Radio, July 1st, 2014, 2:00pm EST

More than 1,000 tornadoes hit the United States and Canada each year, resulting in deaths and property damage, sometimes on a catastrophic scale (National Oceanic and Atmospheric Association (NOAA), 2014b). For example, the May 2011 tornado in Joplin, Missouri, resulted in 159 deaths and caused nearly USD 3 billion in damage (Masters, 2012; NOAA, 2015). In Europe, all but a handful of countries have experienced tornadoes (Groenemeijer & Kuhne, 2014), with many European nations currently improving their severe storm forecasting systems (Rauhala & Schultz, 2009).

Field reports of severe storms are essential to tornado warnings and forecasting because remote sensing data cannot definitively identify some critical storm features (Brotzge & Donner, 2013; Doswell III, Moller, & Brooks, 1998; League et al., 2010; NOAA, 2011). These forecasts are limited by the computational challenges of data processing, insufficient observations of high spatiotemporal resolution, and the lack of valid radar- or satellite-visible predictors for some tornadic weather patterns (Brotzge & Donner, 2013). Moreover, there are too few trained storm spotters to reliably provide warnings with the lead time needed for effective action (Rauhala & Schultz, 2009). As a result, members of the public must often evaluate the risks on their own (Brotzge & Donner, 2013; Doswell III et al., 1998; Durage, Wirasinghe, & Ruwanpura, 2012). Their success depends on how well they can gauge the risks from observable cues (Brotzge & Donner, 2013; Brotzge & Erickson, 2010; Lindell & Perry, 2012). As in the CBC broadcast, valid cues can be both visual (e.g., clouds, wind, hail) and auditory (e.g., the “roar of a ‘freight train’ ”) (Lindell & Perry, 2012, p. 168; NOAA, 2011).

Weather forecasters’ communications attempt to help members of the public take appropriate safety measures. Research has found that people often seek visual and auditory cues before deciding whether to heed them (Dash & Gladwin, 2007; League et al., 2010; Lindell, 2012; Lindell, Lu, & Prater, 2005; Liverman & Wilson, 1981; Sorensen, 2000). For example, the sight of rising water levels has been found to prompt residents to evacuate before an approaching hurricane (Morss & Hayden, 2010). However, relatively little is known about how and how well laypeople extract information from most environmental cues. Here, we offer a general approach to addressing this question, within the normative-descriptive-
prescriptive framework of behavioural decision research (Edwards, 1954; Fischhoff & Kadvany, 2011), illustrated with lay evaluation of visual cues for tornado risk. We begin by briefly summarizing meteorological research into the diagnostic value of those cues (normative analysis). We then assess lay observers’ ability to extract diagnostic information from pictures of clouds, based on methods from decision science (descriptive analysis). These results then frame recommendations for policies designed to accommodate and reduce the limits to lay abilities (prescriptive analysis). In addition to advancing the understanding of tornado risks, we hope to illustrate a method applicable to other environmental hazards, both immediate (e.g., floods) and long-term (e.g., living in an area prone to flooding).

The training of storm spotters is based on meteorologists’ analysis of the cues most useful for predicting severe thunderstorms and tornadoes. US National Weather Service (NWS) training emphasizes visual cues (NOAA, 2011), such as wall clouds, a characteristic lowering of the storm from which tornadoes descend, and having a “solid or hard-looking storm tower with a cauliflower appearance,” indicating strong storm activity (NOAA, 2011, p. 22). Figure 1 shows examples, all taken from official NWS materials or from professional weather photographers, sources that included verification of the weather scene in the photograph (see Methods). Figure A1 in Supplementary Information (SI) A provides further examples. Unfortunately, no cue is perfectly predictive and the most predictive ones, such as the formation of a funnel cloud (Figure 1F), occur so late that they provide little response time. Moreover, some important cues can be difficult to distinguish from less tornadic phenomena (NOAA, 2011; NWS, 2014b). For example, shelf clouds (Figure 1C) may look very similar to wall clouds. Although storm spotters are trained to make these distinctions (NOAA, 2011; NWS, 2014b), even they sometimes have difficulty (Brotzge & Donner, 2013; NOAA, 2011; NWS, 2014b). SI A describes more fully the tornado detection and warning procedures and the performance metrics that provide the normative analysis underlying the present research. In this study, we focus on North America, where tornadoes pose the greatest threat and where tornado forecasting is most developed; see Rauhala and Schultz (2009) for a comparison of the American and European systems.

Here, we present descriptive research into how and how well laypeople detect tornado danger from clouds, applying two approaches from psychophysics, the study of psychological responses to physical stimuli: signal detection theory (Green & Swets, 1988; Macmillan & Creelman, 2004; Wickens, 2001) and multidimensional scaling (Baird & Noma, 1978; Borg, Groenen, & Mair, 2012; Borg & Groenen, 2005). Signal detection theory (SDT) separates decision-maker ability (called sensitivity) to detect signals (here, tornado potential) from the decision rule for acting as though a signal exists. Thus, SDT recognizes that false alarms for tornadoes depend on both how well forecasters can detect them and how cautious forecasters
Example Stimuli

Figure 1. Example stimuli (cloud photographs). A: A supercell thunderstorm with a “cauliflower-like,” hard-textured appearance. B: A wall cloud. C: A shelf cloud. D: Blue skies, with a few cirrus clouds. E and F: Tornadoes, with their funnel clouds clearly visible. Note that D, E, and F form the attention-checking stimuli (see Methods). Photos courtesy of NOAA (A, C, E, F), Roger Edwards (D), and Marko Korošec (B).
want their forecasts to be (Brooks, 2004; Harvey, Hammond, Lusk, & Mross, 1992). Similarly, SDT can distinguish between situations where people fail to seek shelter (a) because they do not see the risk, and (b) because their belief about the odds of a tornado occurring is incorrect or because the inconvenience of a false alarm feels too high. Multidimensional scaling (MDS) extracts the dimensions that decision-makers use in determining the extent to which stimuli are similar (e.g., how much cloud formations look like an archetype of a tornado), and can reveal perceptual processes that individuals may not realize or be able to articulate (e.g., “It just looks like a tornado.”). These methods formalize the analysis of uncertain situations.

We use separate, but interrelated, tasks for SDT and MDS. The SDT task asks participants to judge whether each of 50 pictures of clouds drawn from public sources was taken when a tornado watch was in effect. We elicited a categorical judgment, rather than a continuous one, given the meteorological difficulty of assigning degrees of tornadicity to cloud formations. We chose “tornado watch” as a relatively familiar, and important, category; see the Discussion for more details. The MDS task asks participants which picture seems more tornadic, in pairs involving 11 of the 50 pictures. All participants performed both tasks, with half starting with SDT and half with MDS. We expected that performing the MDS task first would improve SDT performance as a result of providing an opportunity to study some of the stimuli (de Zilva & Mitchell, 2012; Mundy, Honey, & Dwyer, 2007). On the other hand, we did not expect the SDT task to affect MDS judgments, thereby producing an asymmetrical transfer effect, whereby one task affects the other but not vice versa (Poulton & Freeman, 1966).

Our experimental design also compares two probability response modes for the SDT task (Lichtenstein, Fischhoff, & Phillips, 1982). The half-range task has participants decide whether a picture was taken during a tornado watch, and then give their confidence in that judgment with a probability in the [50%–100%] range. The full-range task has participants give a probability judgment for the picture having been taken during a tornado watch, using the [0%–100%] range. We expected the half-range response mode to elicit slightly better sensitivity, by encouraging greater focus on discrimination in its first stage.

Understanding the effects and nuances of tornado experience is a topic of interest in natural hazards research (Blanchard-Boehm & Cook, 2003; Drost, 2013; J. Demuth, personal communication, November 26, 2014; Nagele & Trainor, 2012). As a result, we also collected information about participants’ tornado experience, weather knowledge, emergency preparedness, numeracy, and demographic characteristics, as covariates for modelling

---

1A library of all the stimuli amassed during the creation of this experiment is available at http://goo.gl/WXQPuW. A subset of the library was used in the experiment (as indicated in the library).
Figure 2. A screenshot of a question from the signal detection task (half-range version). Photo courtesy of NOAA.

performance in the SDT task.

**Methods**

Participants \((N = 400)\) recruited via Amazon Mechanical Turk completed three tasks: signal detection theory (SDT) evaluation of cloud tornadicity, multidimensional scaling (MDS) comparison of cloud similarity, and personal information questions. Participants were randomly assigned to one of four cells in a \(2 \times 2\) design, with the factors of (a) SDT response range (half-range or full-range) and (b) task order (SDT or MDS first). Details appear in SI B.

The SDT and MDS tasks use photographs of clouds taken from public sources. Each picture was classified as tornadic or non-tornadic, based on its description by its source and
NWS criteria (NOAA, n.d.-a, n.d.-b). There were 25 pictures of each type.

Figures 2 and 3 show examples of the two SDT tasks, half- and full-range, differing solely in how responses were elicited. The “tornado watch” formulation of both prompts was designed to elicit a judgment of whether a cloud formation was tornadic or non-tornadic, which we then compared with its actual classification based on the NWS criteria.

Participants saw 50 pictures, without being told the 50% base rate of tornadic stimuli. Three of the 50 pictures were exceptionally easy (two Wizard of Oz-type tornadoes, one clear blue sky), in order to assess whether participants were paying attention (see Figure 1). The two tornadoes opened and closed the task. The other 48 stimuli (including the clear blue sky) were in a separate random order for each participant. Response time data were collected in order to see if people spent more time in one condition than another.

After completing these judgments, participants answered three open-ended questions:

- In 25 words or less, what does a “tornado watch” mean to you?
• What made cloud pictures look like they were taken during a tornado watch?
• What made cloud pictures look like they were not taken during a tornado watch?

In the MDS task, participants received 55 pairs of pictures, including all combinations of 11 drawn from the full set. These 11 pictures were chosen to create pairs varying widely in similarity, so that the MDS algorithms could more easily uncover the psychological dimensions underlying participant judgments (Borg et al., 2012); 6 were tornadic and 5 non-tornadic. See Figure 4.

Figure 4. A screenshot of a question from the multidimensional scaling task. Photos courtesy of NOAA.

After completing these tasks participants were asked for their age, gender, education, and state/province of residence. They also answered questions about their tornado experience, interest and knowledge regarding weather, and level of emergency preparedness. Finally, they completed two standard instruments: (a) the 7-item Weather Salience Questionnaire (Short Form), designed to gauge the psychological significance of weather (Stewart, Lazo, Morss, & Demuth, 2012; Stewart, 2009); and (b) the 8-item Subjective Numeracy Scale (Fagerlin et al., 2007), which has been found to predict individuals’ ability to understand risk communications (Zikmund-Fisher, Smith, Ubel, & Fagerlin, 2007). Table 1 provides some example questions.
Table 1
*Example personal information questions. See SI E for the full list of questions.*

<table>
<thead>
<tr>
<th>Topic</th>
<th>Example Items</th>
<th>Response Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather salience</td>
<td>“I take notice of changes that occur in the weather.”</td>
<td>Never (1), Seldom (2), Sometimes (3), Usually (4), All of the time (5).</td>
</tr>
<tr>
<td></td>
<td>“I notice how the clouds look during various kinds of weather.”</td>
<td>Never (1), Seldom (2), Sometimes (3), Usually (4), All of the time (5).</td>
</tr>
<tr>
<td>Subjective numeracy</td>
<td>How good are you at working with fractions?</td>
<td>Six-point scale (1–6), from “Not at All” to “Extremely.”</td>
</tr>
<tr>
<td></td>
<td>How good are you at working with percentages?</td>
<td>Six-point scale (1–6), from “Not at All” to “Extremely.”</td>
</tr>
<tr>
<td>Experience</td>
<td>Have you ever experienced a tornado directly?</td>
<td>Yes (1), No (0).</td>
</tr>
<tr>
<td></td>
<td>Do you have friends or family who have experienced a tornado directly?</td>
<td>Yes (1), No (0).</td>
</tr>
<tr>
<td>Weather interest</td>
<td>How knowledgeable are you about the weather?</td>
<td>Four-point scale (1–4), from “Not at All” to “Extremely.”</td>
</tr>
<tr>
<td></td>
<td>How interested are you in the weather?</td>
<td>Four-point scale (1–4), from “Not at All” to “Extremely.”</td>
</tr>
</tbody>
</table>

**Results**

After describing the sample’s demographic characteristics, we present results for the SDT and MDS tasks. Finally, we analyze individual-level predictors of SDT performance.

**Sample Demographics**

The sample had 156 men and 244 women, whose age ranged from 18–75 years old (*mean* = 37, *median* = 33). Three reported “some high school” as their highest level of education, 48 “high school,” 20 “community college,” 130 “some college,” 143 “college,” 47 “graduate school,” and 9 “professional school.” Participants were widely distributed geographically, allowing for varied tornado experiences (as reported below).

Twenty-two participants failed at least one of the attention-check questions. All analyses were run with and without those individuals, and their exclusion does not change any of the substantive results.
Signal Detection Results

**Sensitivity (d’).** Signal detection parameters were estimated using participants’ choices (i.e., tornadic or non-tornadic) and confidence judgments. In the half-range mode, those responses were used directly. In the full-range mode, 50% was taken as the cutoff (randomly treating 50% responses as tornadic or non-tornadic) and the absolute deviation from 50% was used as the measure of confidence (following Juslin, Olsson, and Bjorkman (1997), who provide a normative analysis of this procedure).  

In order to estimate sensitivity (d’), we first computed the *area under the curve* (AUC) of the receiver operating characteristic (ROC) curve for each participant; see Figure 5. AUC is bounded between 0 and 1. To create an unbounded measure, which can be compared to other methods for estimating $d’$ (see SI C), we also computed $d’ = z(AUC) \cdot \sqrt{2}$, where $z$ is the quantile function of the normal distribution with mean 0 and variance 1. See SI C and Wickens (2001) for details.

Figure 6 shows participant d’s. The sample mean $d’$ is 1.08, with a 95% confidence interval for the mean of (1.04, 1.11). Thus, most participants performed moderately well, with that mean indicating a classification accuracy of about 70% (for individuals trying to maximize the number of correct classifications) (Lynn & Barrett, 2014).

**Bias (c).** The decision criterion (also called the decision bias, decision rule, or response bias) for each participant was estimated from that individual’s hit (H) and false alarm (FA) rates. There are many measures of decision bias. We use

$$c = -0.5 (z(H) + z(FA)).$$

Normatively, $c$ is a function of both the observer’s belief about the base rate of the signal, and their payoff matrix (i.e., their valuation of hits, misses, correct rejections, and false alarms) (Baird & Noma, 1978; Coombs, Dawes, & Tversky, 1970). If participants value the outcomes such that they care only about their classification accuracy, $c$ should reflect their belief about the base rate of the signal, as they should cite the more common category whenever they cannot make a discrimination. Alternatively, participants might differentially value hits, misses, correct rejections, and false alarms, in which case $c$ will reflect that bias.

---

2 This convention allows converting half-range responses to full-range, and vice versa.

3 Changes in the NWS criteria used to classify the stimuli as tornadic or non-tornadic could change based on advances in the meteorological understanding of tornadoes. While such a change would necessarily affect our estimates of the signal detection parameters, it would not impact our conclusions. Any reader planning to use our values as estimates of the population’s psychophysical parameters should be aware of this uncertainty. We thank an anonymous reviewer for highlighting this point.

4 The AUC calculation, used to estimate $d’$, is based on an ROC curve, which shows H and FA rates associated with alternative decision rules. As a result, it provides no measure of bias.

5 In order to account for H or FA equaling 1 or 0, we added 0.5 to each cell of the contingency table. This converts, for example, an H of 1 (i.e., all 25 tornadic photographs correctly classified as tornadic) into the fraction $\frac{25.5}{26}$, and similarly, a FA of 0 into $\frac{0.5}{26}$ (Hautus, 1995; Macmillan & Creelman, 2004).
Figure 5. Each point along the receiver operating characteristic (ROC) curve provides the hit (H) and false alarm (FA) rates for an observer with a decision criterion c, given a fixed value of $d'$. An observer who performs no better (or worse) than chance ($d' = 0$) has the ROC curve given by the line $H = FA$, as seen in the figure. The ROC curve lying above that line depicts the performance of an observer who has performed better than chance ($d' > 0$). The area between the ROC curve and the horizontal axis provides a measure of sensitivity, called the area under the curve (AUC). Empirically, the most common ways of eliciting an ROC curve are by performing multiple experiments, inducing a change to the decision criterion each time, or by using confidence ratings within a single experiment (as in our task). In the latter, one can graph the ROC curve by computing the pair (FA, H) while varying the $x\%$ (in $[0-100\%]$) used as the “tornadic” cutoff. See, e.g., Macmillan and Creelman (2004); Wickens (2001) for more details.
Figure 6. Histograms of participant sensitivities ($d'$) and decision criteria ($c$), and a plot of $d'$ versus $c$. The plot of $d'$ versus $c$ shows that $d'$ and $c$ are roughly uncorrelated ($r = -0.098$).
We chose $c$ as a measure of bias because it has several attractive properties (Macmillan & Creelman, 2004, chap. 2), including its easy interpretation. If $c$ reflects the observer’s valuation of hits, misses, correct rejections, and false alarms, then $c = 0$ (called a “neutral” decision rule) indicates that signal (tornadic) and noise (non-tornadic) stimuli are of equal importance to the observer, positive values (called “conservative”) indicate less tolerance for false alarms relative to misses, and negative values (called “lax”) indicate greater tolerance for false alarms relative to misses (Baird & Noma, 1978; Macmillan & Creelman, 2004; Wickens, 2001). If $c$ solely reflects beliefs about the base rate of the signal, analogous interpretations exist for neutral, conservative, and lax criteria: a neutral criterion implies a perceived base rate of 50%, a conservative criterion implies the perception that the signal rarely occurs, and a lax criterion implies the perception that the signal is common.

The mean $c$ is $-0.02$, with a 95% interval of $(-0.06, 0.02)$. Figure 6 shows a histogram of participant $c$s. On average, participants had an approximately neutral decision criterion, but with a range of individual policies. Figure 6 also contains a plot of $d'$ versus $c$, showing that $d'$ and $c$ are roughly uncorrelated ($r = -0.098$).

According to a strict utility interpretation, the positive $c$ values among the distribution of $c$s in Figure 6 would indicate that many participants wanted to avoid false alarms at the cost of increasing the number of misses, which seems unlikely given the dangers posed by tornados. Instead, we think the diversity of $c$ values observed could be better explained by participants attempting to maximize their accuracy, setting their $c$s to reflect their beliefs in the odds of a tornadic stimulus occurring in the task. Similarly, it is likely that the nature of the task – asking for a tornadicity classification, rather than, say, about what the participant would do given the weather scene – supports a probability interpretation over a utility interpretation. Thus, we shall interpret $c$ as reflecting participants’ priors, recognizing that some participants may also have incorporated utilities into their decision criteria.

**Stimulus-Level Results.** Figure 7 shows how accurately the individual stimuli were classified as tornadic (signal) or non-tornadic (noise), based on the mean full-range probability assigned (using all responses, converting those given for the half-range task to probabilities on the 0–100% scale; see footnote 2). As can be seen, although the mean probability judgment was higher than 50% for most tornadic clouds (labelled with an “s”) and lower for most non-tornadic ones (labelled with an “n”), there were some exceptions.\(^6\)

The MDS scaling results guided our interpretation of the features affecting the accuracy of

\(^6\)Note that, in Figure 7, some of the tornadic stimuli below 50% (s6, s26, s35) are very close to the 50% mark, as is the dust storm (n2) that is above 50%. Similarly, a supercell (s40) and a wave cloud (n13) are very close to the 50% mark, though these are above and below 50%, respectively. In fact, if we consider a stimulus misclassified if the 95% confidence interval for its mean rating contains 50%, then s40 and n13 are misclassified as well.
these judgments.

**Multidimensional Scaling**

MDS creates a geometric representation of the stimuli using participants’ judgments of the more tornadic picture among each of the 55 paired comparisons created from the 11 selected pictures. Using standard criteria (see SI D), we found the best fit with the two-dimensional representation in Figure 8, with dimensions that we interpret as reflecting (a) the darkness or “ominousness” of the picture and (b) its clarity, referring to how easy it is to discern and interpret the features of the cloud formation that mark it as tornadic or non-tornadic. Looking at the specific pictures, we find potential sources for misclassifications: pictures of (non-tornadic) shelf clouds (1, 3) and (tornadic) supercells (6, 7, 8) were seen as similar. Pictures with upper- and mid-level tornadic cloud features (9, 10) were seen as similar to non-tornadic fair weather cumulus clouds (2). Among the tornadic photographs, 9 and 10 were seen as the most dissimilar to the tornadic extreme represented by 11. A unidimensional scaling (provided in SI D) confirmed the upper- and mid-level tornadic clouds are likely to be confused with non-tornadic clouds. Put simply, shelf clouds look worse than they are because they are both ominous and dark, with structural features that might seem unfamiliar to the untrained eye. Clouds with upper- and mid-level tornadic features are slightly unusual, but the blue skies that surround them mean they are fairly bright, and (we speculate) less ominous to the layperson. These results were not affected by whether participants performed the MDS task first or the SDT task first.

The MDS results shed light on the systematic misclassifications that occurred in the SDT task (Figure 7). All of the shelf clouds (n3, n19, n20, n27, n29) were classified inaccurately, as were four of the five with upper- and mid-level cloud features (s2, s6, s26, s27). The other tornadic pictures that were misclassified contained *mammatus* clouds (s24, s35): the MDS would suggest this was because of their brightness. A picture of lightning (n31) and a dust storm (n2) were misclassified as tornadic: the former is certainly ominous, and the latter is dark, with features that are (almost “by definition”) difficult to discern.

We directly tested the link between the MDS results and SDT results by regressing the mean SDT full-range scores of the 11 common stimuli onto their coordinates in the two-dimensional MDS space. Table 2 displays the estimated coefficients for the darkness and clarity dimensions, which account for a large proportion of the variance ($R^2 = 0.97$). We see that the darkness/ominousness dimension is clearly the more important dimension, and that the linear regression does a good job of modelling the SDT scores, as would be expected if MDS captures the attributes underlying judgments of tornadicity. We also regressed each individual’s SDT scores on the MDS coordinates, to see how well individual SDT scores were
Figure 7. Boxplots of sample means for all stimuli ratings on the full-range, 0–100% scale, corresponding to the judged “Probability this picture was taken during a tornado watch.” An “s” specifies a tornadic (i.e., “signal”) stimulus, and “n” a non-tornadic (i.e., “noise”) stimulus. The labelling scheme corresponds to the stimulus library (see footnote 1). The horizontal line is placed at 50% to more easily separate those stimuli with mean ratings above/below 50% (see the main text).
Two-Dimensional Scaling

*Figure 8.* Two-dimensional scaling of the 11 stimuli in the MDS task. The horizontal dimension refers to the darkness/ominousness of the weather scene. The vertical dimension refers to the clarity of the scene, i.e., how easy it is to discern and interpret the important structural features of the cloud formations. Note that, with respect to the stimulus library (see footnote 1), stimulus 1 is n3, 2 is n15, 3 is n19, 4 is n31, 5 is n22, 6 is s11, 7 is s21, 8 is s23, 9 is s26, 10 is s27, and 11 is s31. Photos courtesy of NOAA (1, 2, 3, 6, 7, 8, 9, 10, 11) and Roger Edwards (4, 5). Stimuli 3, 6, and 8 have been shifted from their true positions to reduce overlap: SI D contains the same scaling without the overlay of photographs.
modelled by the aggregate MDS solution. The average $R^2$ was 0.7. Note that to make interpretation easier, the coordinates of the MDS representation were standardized, and the space transformed so that its origin represents the location of a hypothetical stimulus having essentially no “darkness/ominousness” and being the least “clear” possible.

Table 2

<table>
<thead>
<tr>
<th>Coefficient estimates and standard errors for a regression modelling the mean SDT full-range scores (i.e., mean judged $\text{prob(tornadic)}$) as a function of the MDS coordinates.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable:</strong></td>
</tr>
<tr>
<td>$\text{darkness/ominousness}$</td>
</tr>
<tr>
<td>(1.57)</td>
</tr>
<tr>
<td>$\text{clarity}$</td>
</tr>
<tr>
<td>(1.57)</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>(4.01)</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>$R^2$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
</tr>
<tr>
<td>Residual Std. Error</td>
</tr>
<tr>
<td>F Statistic</td>
</tr>
</tbody>
</table>

**Note:** *p<0.1; **p<0.05; ***p<0.01

Individual Difference Measures

Participants’ responses to the seven questions of the Weather Salience Questionnaire (Short Form) were summed (with question 5 reverse scored) to form a weather salience score (WxSQ), which had a Cronbach’s $\alpha$ of 0.61. Participants’ responses on the subjective numeracy questions were averaged to produce a subjective numeracy score (SNS), which had a Cronbach’s $\alpha$ of 0.86. We created a Tornado or Dixie Alley score (0, 1), reflecting whether participants were in one of two tornado-prone regions. The former includes Texas, Oklahoma, Kansas, Missouri, Iowa, Nebraska and South Dakota ($n = 54$) and the latter Louisiana, Arkansas, Mississippi, Alabama, Tennessee, and Georgia ($n = 35$). The three questions regarding knowledge, interest, and forecast-following behaviour were transformed to a 1–4 scale, and then averaged (and standardized) to create a (self-reported) meteorophily.

---

7 Cronbach’s $\alpha$ is a measure of internal consistency (Cronbach, 1951). Bland and Altman (1997) provide a short and accessible introduction to Cronbach’s $\alpha$.

8 Though there are no formal definitions for these areas, we used those from NOAA: see [http://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology/trends](http://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology/trends).
score, which had a Cronbach’s $\alpha$ of 0.68. The nine items related to participants’ severe-storm experience were summed (and then standardized) to create an experience score, which had a Cronbach’s $\alpha$ of 0.76. The item measuring the impact of participants’ tornado experience was kept separate and standardized. Having emergency supplies, a planned place to take shelter, and knowing that a tornado warning is more serious than a watch—all binary variables—were kept separate and untransformed. Finally, we created a binary variable fail, equal to 1 if the participant failed any of our attention checks, and 0 otherwise. SI E further describes the creation of these measures.

Modelling SDT Parameters

Table 3 predicts $d'$ and $c$ as the dependent variables, using the individual difference measures and the experimental condition factor variables (SDT response mode and task order) as covariates. See SI F for regression diagnostics.

The most striking results are the higher values of $d'$ for participants who report having had more personal experience with tornadoes and having done more to prepare for them, and the lower values of $d'$ among those who report greater impacts from tornadoes and those who failed the attention checks. None of the following variables predicted greater sensitivity ($d'$): living in Tornado or Dixie Alley, finding weather more salient (WxSQ), reporting greater meteorophilia, or knowing that a tornado warning is more serious than a tornado watch. Decision bias ($c$) was related to two individual difference measures: participants with higher WxSQ scores (indicating that weather was more salient for them) were more likely to see danger in the cloud formations, as were participants who reported having planned a place to shelter during severe weather.

Because of the interaction in the model, the coefficients on order and mode in Table 3 show simple effects (i.e., the effect of task order and of response mode when the other variable is set to 0). We used the results in Table 3 and the variance-covariance matrix of the coefficients to investigate main effects as well. We found that $d'$ was unrelated to whether the SDT task came first or second, and whether the SDT task used the half- or full-range mode. However, participants were significantly less likely to categorize stimuli as tornadic (i.e., had higher $c$ values) when they completed the SDT task first, and significantly more likely to categorize stimuli as tornadic (i.e., had lower $c$ values) when responding using the half-range mode. The estimates for the main effect of task order and response mode on $c$ were 0.12 and $-0.11$, respectively, with 95% confidence intervals of $(0.04, 0.20)$ and $(-0.19, -0.03)$.

A secondary analysis found that living in Tornado Alley predicts a 0.1 increase in $d'$, while living in Dixie Alley predicts a 0.1 decrease in $d'$. The two groups combined, however, had only 89 people, from the sample of 400.

We thank an anonymous reviewer for suggesting this probability-based interpretation.
Table 3  
Coefficient estimates and standard errors for regressions modelling \( d' \) and \( c \), based on the individual difference measures and the experimental conditions.

<table>
<thead>
<tr>
<th></th>
<th>( d' )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>order (0 = MDS-SDT)</td>
<td>-0.02</td>
<td>0.17***</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>mode (0 = full-range)</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>experience</td>
<td>0.07***</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>impacts</td>
<td>-0.05***</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>alley (1 = lives in Tornado/Dixie Alley)</td>
<td>0.02</td>
<td>-0.0002</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>meteorophily</td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>shelter (1 = has planned shelter)</td>
<td>0.08**</td>
<td>-0.10**</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>supplies (1 = has supplies)</td>
<td>-0.05*</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>watch-warning (1 = knows warning is more serious)</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>WxSQ (weather salience)</td>
<td>0.01</td>
<td>-0.02***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>SNS (Subjective Numeracy Score)</td>
<td>0.03**</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>fail (1 = failed attention checks)</td>
<td>-0.45***</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>order ( \times ) mode</td>
<td>0.03</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.75***</td>
<td>0.63***</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.20)</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>R(^2)</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td>Adjusted R(^2)</td>
<td>0.17</td>
<td>0.07</td>
</tr>
<tr>
<td>Residual Std. Error (df = 386)</td>
<td>0.30</td>
<td>0.41</td>
</tr>
<tr>
<td>F Statistic (df = 13; 386)</td>
<td>7.39***</td>
<td>3.39***</td>
</tr>
</tbody>
</table>

**Note:** *p<0.1; **p<0.05; ***p<0.01
Discussion

Psychophysical Performance

The present study provides the first psychophysical investigation of tornado risk perception. Our signal detection analysis found that participants had some ability to distinguish between tornadic and non-tornadic stimuli, with the mean $d'$ of 1.08 equating to about 70% accuracy for individuals with the roughly neutral decision rule found here ($c = -0.02$). The low correlation between $d'$ and $c$ (Figure 6) shows that participants’ decision criteria are independent of their perceptual ability to detect tornadic weather.

With respect to individual differences in psychophysical performance (Table 3), participants who reported more experience with tornadoes demonstrated greater discrimination ability, whereas those who reported greater impacts from tornadoes demonstrated less and living in a tornado-prone area had no effect. However, $d'$ was higher for individuals who reported having planned a place to shelter, who rated themselves higher on the SNS, and who passed our attention checks. Decision bias ($c$) was associated with only two individual difference measures. Participants with a higher weather salience score (WxSQ) were more likely to categorize cloud formations as tornadic, as were those who reported having planned a place to shelter.

Discrimination was the same with the half-range and full-range SDT tasks, contrary to our prediction that the comparison required by the half-range task would lead people to look more closely. Participants did, indeed, spend an average 1.25 seconds longer on it per judgment; however, that time apparently went into completing the task, rather than thinking about its content. For $c$, participants using the half-range mode had a more lax criterion (i.e., were more likely to categorize a stimulus as tornadic). We had no prediction regarding the effect of response mode on $c$, but speculate that explicitly choosing the category in the half-range mode made utilities more salient than in the full-range task, where the category was inferred from the confidence rating. As a separate methodological point, the best single predictor of $d'$ was the attention-check indicator variable. Participants who improperly classified any one of the two funnel clouds or the clear blue sky showed poorer discrimination ability on the other 47 stimuli, supporting the utility of such checks (Downs, Holbrook, Sheng, & Cranor, 2010; Oppenheimer, Meyvis, & Davidenko, 2009).

We also found no better discrimination among participants who did the MDS task first, contrary to our prediction that examining a subset of the stimuli would improve performance. However, $d'$ calculated for just the 11 stimuli used in the MDS task was higher for those who did the MDS task first (mean of 0.63 versus 0.56, although the 95% confidence interval for the difference in means, $(-0.02, 0.17)$, includes no difference). For related results, see de
Zilva and Mitchell (2012); Lavis and Mitchell (2006); Mundy et al. (2007), who investigate the features of prior exposure that improve discrimination of stimuli and memory for their attributes. Participants who performed the MDS task first were more likely to classify stimuli as tornadic in the SDT task (had lower c values), suggesting that prior exposure to stimuli led to an increased sense of tornadic risk.

The MDS results suggest heuristics that might affect performance—in the SDT task and perhaps in life. Specifically, associating darkness with tornado risk may lead to viewing upper- and mid-level tornadic clouds (the visible blue sky) as less tornadic than they are; such clouds require some sophistication to notice and understand the features that mark them as dangerous, the lack of which could also lead to misclassification. Regression showed that the darkness/ominousness and clarity dimensions effectively modelled the mean response in the SDT task for the 11 stimuli included in both the MDS and SDT tasks. It also showed that darkness/ominousness was much more important than clarity; similarly, additional MDS analyses (see SI D) found that the darkness/ominousness dimension was more important than clarity in explaining judgments of cloud similarity. As cues of actual (rather than perceived) tornadicity, darkness/ominousness has validity, given that severe storms are dark when they are overhead or very close by. On the other hand, cloud formations with more easily discernible features (i.e., higher on the clarity dimension) are, to the best of our knowledge, no more or less likely to be tornadic.11 Future work could, by scaling a large set of stimuli, clarify both the link between the dimensions laypeople are using and their validity as markers of tornadic threat, and the difficult-to-define “clarity” dimension.

Our approach classified stimuli dichotomously, as tornadic or non-tornadic. An alternative approach is to characterize the stimuli by their probability of resulting in a tornado. However, we could not find those probabilities in the literature, nor the observational data with which to construct them ourselves. By using the NWS storm-spotter training curriculum as a classification guide, we have relied on expert opinion to decide what is tornadic.

Participants were prompted to classify the stimuli with a question that asked about the occurrence of a tornado watch (Figures 2 and 3). We considered many alternatives to that chosen formulation, including: “Are these clouds associated with tornadoes?”, “Are these clouds often associated with tornadoes?”, “Will these

11 The validity of the dimensions can also be investigated by regressing the actual classifications (tornadic or non-tornadic) of the 11 stimuli in the MDS task on their coordinates in the MDS space. This logistic regression finds that, for a 1-unit increase on the darkness dimension, the odds of a stimulus being tornadic are 2.2 times greater. (The coordinates have been standardized, thus 1-unit corresponds to one standard deviation.) For a 1-unit increase on the clarity dimension, the odds of a stimulus being tornadic are 1.1 times greater. However, these parameters cannot be precisely estimated due to the small sample size.
clouds form a tornado?”, with appropriate changes to other instructions. While recognizing that some (more sophisticated) participants might have preferred a less deterministic formulation, we saw no tractable way of expressing “associated” so that it would be interpreted similarly by all respondents (Wallsten, Fillenbaum, & Cox, 1986). One concern with our chosen “watch” formulation was whether participants would correctly interpret the prompt. Notwithstanding possible concerns about self-reports (Ericsson & Simon, 1980), pre-testing found that participants generally interpreted it as desired, and the open-ended questions asked after the SDT task found that participants in the actual experiment generally did so as well.

Policy Implications

Our results suggest that lay observers have generally effective heuristics for deciding whether clouds are tornadic. However, those heuristics lead to biases, including systematically misclassifying shelf clouds as tornadic, clouds with upper- and mid-level tornadic features (such as well-defined storm towers) as non-tornadic, and mammatus clouds as non-tornadic. Looking at the MDS representation (Figure 8), projecting onto the horizontal dimension produces a mostly correct “continuum” from non-tornadic to tornadic phenomena, but with the shelf-clouds intermixed among the tornadic clouds, and the MDS results further showing that clouds with upper- and mid-level tornadic features are likely to be confused with non-tornadic clouds.

Even though the precise $d'$ and $c$ values observed here cannot be confidently extrapolated to a real world setting, the patterns of results have implications for hazard training, creating risk communications, and anticipating inappropriate responses to environmental cues. Specifically, they show that lay observers have acquired some ability to assess tornadic potential, which could be improved were it possible to remedy their biases. The results also provide quantitative measures of behaviour for use in evaluating policy options (e.g., alternative warning strategies).

Our results suggest that training should take advantage of the attributes that emerged as most salient, namely the “darkness” and “clarity” of cloud scenes. It should also emphasize the difference between shelf clouds and more tornadic formations (e.g., wall clouds), while devoting extra attention to upper- and mid-level tornadic cloud features as well as mammatus clouds. Testing the possibility of such training is a logical next step in the research, which might be pursued not only with lay observers, but also with storm spotters, especially in light of the increased interest in their performance (Jans, Allison, & Denzer, 2011; Jans & Keen, 2012; Klenow & Reibestein, 2014). Deficiencies in storm spotters’ $d'$s would indicate the need for improved training to increase perceptual acuity; MDS results
could suggest how to direct that training. As an anonymous reviewer remarked, forecasters would have difficulty interpreting the reports of storm spotters with biased decision criteria, especially if they vary across spotters, making the implementation of NWS policy more difficult.

Better discrimination can also reduce a natural concern of forecasters: having false alarm rates so high that people tune out their warnings (a “cry wolf” effect) (Barnes, Gruntfest, Hayden, Schultz, & Benight, 2007; Brooks, 2004; Brotzge & Donner, 2013; Durage et al., 2012; League et al., 2010; Ripberger et al., 2014; Simmons & Sutter, 2009). Since 2007, the NWS has used warning polygons in order to make its warnings more specific to the areas under threat, in contrast to the broader, county-wide warnings still practiced in Canada or the heterogeneous European warning strategies (Durage et al., 2012; Environment Canada, 2014; NWS, 2008; Rauhala & Schultz, 2009). Our results suggest an alternative strategy: tailor tornado warnings to the cues available to the public, based on proximity to the storm system. For example, individuals not directly under a tornadic storm might hear a warning, but only have upper- or mid-level cues available to them. A forecast could be more useful to these individuals if it provided grounds for concern regarding features that forecasters can see, but these individuals cannot, such as the storm’s movement and radar-visible signs of tornado threat (e.g., a hook echo) (Markowski, 2002). In addition to providing people more time to prepare for storms, that additional information could also help them to judge forecasters more fairly (see also Ripberger et al. (2014)). Of course, the relevance of our results for creating interventions to improve protective action decision-making and communications depends upon the extent to which the public relies on physical environmental cues, which is likely a function of demographic variables, experience, and locale, and an empirical question that should be addressed in future work.

Conclusion

The novel application of psychophysics presented here provides a proof-of-concept of an emerging approach to risk problems: that of understanding human physical perception (Agdas, Webster, & Masters, 2012). These methods could be applied to cues regarding both immediate risks (e.g., flash floods, influenza, Ebola) and long-term ones (e.g., living in a tornado- or hurricane-prone region, or near a nuclear reactor). Understanding what counts as “signal” could improve the evaluation and effectiveness of risk communications (Fischhoff, 2013; Morris & Hayden, 2010; Ripberger et al., 2014), as well as our understanding of processes such as how perception of climate change is influenced by experience of extreme weather events (Akerlof, Maibach, Fitzgerald, Cedeno, & Neuman, 2013; Broomell, Budescu, & Por, 2015; Lefevre et al., 2015; Leiserowitz, 2006; van der Linden, 2015).


Acknowledgements. BD, BF, SB, and AD designed the experiment. BD collected the data, and BD, BF, SB, and AD conducted the analyses. BD and BF wrote the manuscript. The research was funded by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship (awarded to BD) and US National Science Foundation Grant SES-0949710. The authors thank Roger Edwards and Marko Korošec for their permission to use their photographs, and Julie Demuth (National Center for Atmospheric Research), Gabrielle Wong-Parodi (Carnegie Mellon University), Michael Yu (Carnegie Mellon University), Coreen Farris (RAND Corporation), Roger Edwards, and three anonymous referees for their invaluable feedback.
References


Borg, I., Groenen, P. J. F., & Mair, P. (2012). *Applied multidimensional scaling*. Springer. 4, 9, 40, 44, 45, 47, 48


ENVIRONMENTAL RISK PERCEPTION FROM VISUAL CUES

10.1073/pnas.1317504111


ENVIRONMENTAL RISK PERCEPTION FROM VISUAL CUES


NOAA. (2011). Weather Spotter’s Field Guide. 3, 4


