Research Article

Probabilistic Interval Forecasts: An Individual Differences Approach to Understanding Forecast Communication

Margaret A. Grounds,¹ Susan Joslyn,¹ and Kyoko Otsuka²

¹University of Washington (UW), P.O. Box 351525, Seattle, WA 98195-1525, USA

²Tokyo Institute of Technology, Tokyo, Japan

Correspondence should be addressed to Susan Joslyn; susanj@uw.edu

Received 9 February 2017; Revised 15 May 2017; Accepted 13 June 2017; Published 24 September 2017

Academic Editor: Guillermo Baigorria

Copyright © 2017 Margaret A. Grounds et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Predictive interval forecasts, showing a range of values with specified probability, have the potential to improve decisions compared to point estimates. The research reported here demonstrates that this advantage extends from college undergraduates to a wide user group and does *not* depend on education. In two experiments, participants made decisions based on predictive intervals or point estimates and answered questions about them. In Experiment 1, they also completed numeracy and working memory span tests. Those using predictive intervals were better able to identify situations requiring precautionary action. Nonetheless, two errors were noted: (1) misinterpreting predictive intervals as diurnal fluctuation (deterministic construal errors) and (2) judging the probability of events within and beyond the interval, when asked about them separately, as greater than 100%. These errors were only partially explained by WMS and numeracy. Importantly, omitting visualizations eliminated deterministic construal errors and overestimation of percent chance was not consistently related to decision quality. Thus, there may be important benefits to predictive interval forecasts that are not dependent on a full understanding of the theoretical principles underlying them or an advanced education, making them appropriate for a broad range of users with diverse backgrounds, weather concerns, and risk tolerances.

1. Introduction

People make myriad important decisions that involve weather, many related to safety. Reliable uncertainty information is often available, although experts tend to withhold quantitative uncertainty information because they are concerned that end-users will not be able to understand it. Indeed, there is a large literature suggesting that people have trouble understanding probabilistic information [1–5], violate axioms of utility theory, respond to probabilities nonlinearly, and ignore prior probabilities [6–10]. Nonetheless, recent psychological research suggests that people can benefit from fairly complex expressions of uncertainty in the sense that they make better decisions when uncertainty information is available compared to when it is omitted [11].

One example is the predictive interval forecast, which describes a range of values within which the observed value is expected with specified probability, such as 80%. Predictive intervals have the potential to address a variety of

weather concerns. For instance, one user may need to take precautions when nighttime low temperatures are expected to fall below freezing (32°F, 0°C) while another may need to take precautions when temperatures are expected to fall below 25°F (−3.9°C). In theory, a range of values and the associated probability provide useful uncertainty information to both. Indeed, research with college student participants suggests that predictive intervals benefit users in just that way, allowing them to better distinguish, than do point estimates, situations in which to take precautionary action [12, 13]. In these studies, some participants based their decisions on a forecast that included the upper and lower bound temperatures for an 80% predictive interval as well as a median value described at the “most likely” temperature. Others used only the most likely temperature, a point estimate. Both groups made precautionary decisions such as whether or not to post a freeze warning. Those using predictive intervals were more likely to take precautionary action when the threshold of concern, such as 32°F (0°C), was within the predictive interval

and more likely to take nonprecautionary action when the threshold was beyond the interval than were those using the point estimate. In addition, predictive intervals narrowed the range of temperatures participants regarded as likely, suggesting that they function to refine users' understanding of the likelihood of critical outcomes.

However, it is important to note that this uncertainty expression is somewhat complex, describing the probabilities of ranges of values within and outside the interval. Understanding predictive intervals may be dependent on minimal cognitive abilities such as numeracy, the extent to which people understand basic mathematical and probability concepts [14]. It may also be related to working memory capacity (WMC), one's ability to control attention and maintain information in consciousness [15]. Understanding predictive intervals may also be dependent on exposure to prerequisite concepts that would be provided in some minimum level of formal education. The research reported here will explore these issues. If predictive intervals require superior cognitive abilities or advanced education to fully comprehend them, serious misunderstandings could arise in a wider user group that may negate the previously observed benefits.

Indeed, two specific interpretation errors were noted in the earlier research among a subset of college student participants that may increase in a broader population [12]. The first occurred when visualizations were added (a bracket outlining the 80% range; see Figure 1(a)). Some participants mistook the value at the upper bound of the range, located at the top of the bracket, as the most likely daytime high temperature and the lower bound value, located at the bottom of the bracket, as the most likely nighttime low. Participants apparently thought the interval was a group of point estimates for diurnal fluctuation over the 12-hour period. This misinterpretation was referred to as the "deterministic construal error" (DCE) because it transformed what was intended as an uncertainty forecast into a point estimate implying a deterministic outcome. Moreover, those who committed DCEs made significantly different decisions than did those with the correct interpretation, confirming that the misinterpretation concerned the forecast itself [12].

Psychologically similar errors occur with other expressions of forecast uncertainty. The Cone of Uncertainty, for instance, meant to convey the possible path of a hurricane, is often misinterpreted as the wind field or the extent of the storm [16]. The probability of precipitation, meant to convey the percent chance of precipitation, is often misinterpreted as the percent of time or area affected by precipitation [17, 18]. A similar error is noted in other domains. For example, people who were told that the drug Prozac has a 30–50% chance of sexual side effects thought that the percent described the proportion of sexual encounters for the individual rather than the percent chance for the individual [19]. These misinterpretations may arise because they reduce processing load. In the case of the DCE, for instance, the deterministic interpretation, implying a single outcome with certainty, replaces the more complex but accurate interpretation involving multiple possible outcomes and their associated probabilities. In other words, there may be a natural propensity to make this error that can only be overcome by applying logical reasoning skills.

Therefore, overcoming the error may be related to cognitive control, numeracy, or level of education. For that reason, the work reported here examined the relationship between DCEs and these variables.

However, it may be possible to reduce DCEs by using an expression of the predictive interval that better conveys the probabilistic nature of the information, precluding the need for special skills or extra effort. Previous studies [12] tested the bracket shown in Figure 1(a), as well as visualizations specifically recommended for indicating uncertainty [20, 21] such as dotted rather than solid lines and fuzziness. However, similar DCE rates were noted with each expression. Here, we take a different tack and test expressions designed instead to block the misinterpretation. In addition, we ask whether visualizations in general are at fault. It is possible that, motivated by their bias toward certainty, people automatically interpret visualizations as portraying deterministic outcomes, without bothering to fully process accompanying explanatory information. Indeed, there is new evidence that an array of widely accepted techniques for communicating uncertainty (e.g., fuzziness, sketchiness) fail to communicate the notion of uncertainty *at all* to the majority of users [22].

It is important to note that although a substantial minority committed DCEs, most participants in our previous studies correctly interpreted predictive intervals as probabilistic, although they did not interpret it exactly as intended. About half of the participants indicated chances of observations falling within and beyond the interval, asked in three separate questions, which summed to more than 100%. However, surprisingly, this error had no impact on participants' decisions. Here, we explore two possible explanations. In previous studies, questions referring to the range above and below the interval were asked first. Participants tended to indicate values larger than 10% in each case, either because they misunderstood the interval or because they thought it was too narrow. Perhaps they simply failed to correct for this overestimation when indicating the chances that the observation would fall within the interval, which was requested last. Thus, the error may have been at least partially due to question order, held constant in previous studies.

However, there may be more to it than that. This error may be similar to one observed in previous research [23], referred to as "subadditivity," in which participants indicate that the sum of the whole is less than the sum of the parts [24]. For example, participants estimated a 58% chance of death by "natural causes." However, when answering a series of questions targeting subsets of that category, the total summed to 73% chance. Unpacking (e.g., asking questions about each individual cause) may remind people of possibilities that they would not have otherwise considered or it may increase the salience of those possibilities [23, 25]. Perhaps similar explanations account for overestimates for the likelihood of outcomes within and beyond the predictive interval.

It is important to note that how ever the error arises, to make it, one must ignore the logical principle requiring that the sum of the parts equal that of the whole, in this case 100% of the outcomes. Therefore, a possible solution is an expression of likelihood that makes the reference class or whole more salient [23], such as frequency (1 in 10 days like this).

FIGURE 1: Visualizations used in the temperature decision task in Experiment 1. Font sizes are slightly smaller in this figure than they were in the experiment to conform to journal requirements.

Thus, if overestimation errors are due in part to users losing track of the fact that each segment is part of a whole, frequency expressions may reduce these errors because they make “the whole” more salient. Indeed, frequency has been shown to be easier for people to understand in several contexts [26, 27].

Finally, we wondered what participants thought about the distribution of values within the interval. This is an important question because participants’ understanding of the distribution could influence their interpretations of the likelihood of specific values within the interval. Users may regard values as uniformly distributed across the interval or

normally distributed around the most likely value, for which there is some previous evidence in research testing slightly different interval expression [28, 29].

In sum, although the predictive interval is a promising form of uncertainty communication potentially relevant for a wide range of user concerns, misinterpretations noted in previous research suggest that some participants fundamentally and others partially misunderstand the forecast. The proportion of errors may increase in a more diverse population and potentially prevent many users from reaping benefits. Thus, it is important to better understand the cognitive processes that underlie people’s interpretation of predictive intervals,

in particular whether it is related to WMC, numeracy, education, or the manner in which the forecast is expressed. We address these key issues in two experiments.

The primary goal of the first experiment was to examine the role of individual differences in understanding uncertainty expressions so it included measures of WMC and numeracy. Errors such as DCEs or overestimation of the likelihood of observations within and beyond the interval may be due to an inability to maintain all of the relevant information in the focus of attention simultaneously, that is, WMC [15]. In general, human WMC is limited but can vary from person to person. Indeed, evidence suggests that those with low WMC are less likely to provide normative or “rational” responses in classic reasoning tasks [30, 31]. Similarly, limitations in WMC may prevent one from overriding a natural preference for the deterministic interpretation or noticing that the sum of the parts cannot be greater than the whole (subadditivity) for which there is some evidence [32]. Misinterpretations may also be due to low numeracy, the extent to which people understand basic mathematical concepts [14]. Those with low numeracy may not properly appreciate the probabilistic nature of the information provided in the predictive interval leading to DCEs or likelihood estimation errors. Thus, Experiment 1 examined the relationship between participants’ scores on tests of WMC or numeracy and error rates.

The primary goal of the second experiment was to ask whether previously observed benefits, such as better decisions and more precise likelihood estimates, are observed across levels of education. To this end, participants were recruited on the World Wide Web. In both experiments, participants performed the same three weather-related decision tasks. They were required to make precautionary decisions based on temperature forecasts and to indicate the temperature values and likelihoods that they expected. Both experiments tested whether misunderstandings could be reduced by communication methods that targeted those specific errors.

2. Experiment 1

Experiment 1 tested the relationship between cognitive capacities (WMC and numeracy) and misinterpretations of predictive interval forecasts. Participants completed three forecast-based decision tasks as well as a test of WMC and a test of numeracy skills. Experiment 1 also sought to determine whether the orientation of the bracket graphic affected DCEs. Two orientations were tested. In previous studies, the upper and lower bound values were on the left side and the most likely temperature was on the right (see Figure 1(a)). Assuming that people read from left to right, they encountered the upper and lower bound values first, perhaps making them seem more important, that is, the daytime high and the nighttime low temperature. In order to determine whether this was so, a mirror image bracket was also tested, with the upper and lower bound on the right side, to determine whether it reduced errors (see Figure 1(b)).

2.1. Method

2.1.1. Participants. The participants were 644 undergraduates from a large northwestern university in the USA. The mean age was 19.66 and 404 (63%) were females. They were enrolled in the introductory psychology course and received course credit for participating. The majority, 72%, indicated that they preferred the Fahrenheit temperature scale.

2.1.2. Procedure. After providing informed consent, each participant performed the tasks in one of three fixed orders, (1) temperature decision task, numeracy test, working memory test; (2) temperature decision task, working memory test, numeracy test; and (3) working memory test, temperature decision task, numeracy test. The numeracy test was not administered before the temperature decision task because the questions, many about probability, could affect participants’ responses to the temperature decision task.

2.1.3. Decision Task. Participants made three weather-related decision tasks in a fixed order. Each task was accompanied by a two-day forecast showing the daytime high and nighttime low temperature for consecutive days (see Figures 1(a)–1(c)).

(1) In the first task, participants decided whether to issue a freeze warning for an agricultural community to inform farmers when to protect crops from freezing temperatures. According to the instructions, temperatures at or below 32°F (0°C) could cause “destruction of vegetation and potential loss of crop yield.” To provide a realistic cost-loss structure, participants were cautioned against posting the warning when freezing temperatures were not expected because crop protection involved material and labor costs.

(2) In the second task, participants decided whether to issue a “hard freeze warning” also for crop protection, with a 25°F (−3.9°C) threshold temperature. They were told that temperatures below 25°F resulted in “extensive destruction of vegetation and loss of crop yield.” Participants were cautioned against posting the warning when temperatures below 25°F were not expected because protection involved material, fertilizer, and labor costs.

(3) In the third task, participants decided whether to take an elderly relative outside for a walk. According to the instructions, this decision concerned heat-related health consequences with a threshold of 100°F (37.8°C). Participants were told that exposure to temperatures of 100°F or greater could result in “damage to the brain, kidney, or heart.” This task is particularly interesting because, unlike the previous two, *inaction* was the more cautious choice (see Table 1 for a full list of the temperature values).

Each task began with a brief instruction describing task goals and the cost-loss structure. Notice that, in all three tasks, the costs and potential losses are not precisely quantifiable, as is the case in many real-world weather situations. For this reason, there is no economically correct response. After the instructions, participants saw the forecast (Figure 1), which remained on the screen while they answered a series of questions about it. Participants indicated the temperature they expected to observe for the decision-critical time periods (nighttime low for tasks 1 and 2, daytime high for task 3)

TABLE 1: Forecast temperature values in degrees Fahrenheit (°F).

	Friday daytime	Friday nighttime	Saturday daytime	Saturday nighttime
Freeze warning (threshold = 32°F)				
Upper bound	44	36	44	39
Median	41	33	39	36
Lower bound	38	31	34	33
Hard freeze warning (threshold = 25°F)				
Upper bound	40	29	39	32
Median	37	26	35	29
Lower bound	34	23	31	26
Heat health (threshold = 100°F)				
Upper bound	102	71	99	69
Median	99	67	96	66
Lower bound	96	63	93	63

and made a decision for each date. To assess their understanding of the predictive interval forecast, they were also asked to indicate the temperatures they expected to observe for the other two time periods and the probability (11-point horizontal scale from 0 to 100% with response options at 10% increments) that the observed temperature for Day 1 daytime high and nighttime low would be (a) at or above the temperature at the upper bound, (b) at or below the temperature at the lower bound, and (c) between the two bounds (see Figure 2). Only temperature values were mentioned in these questions; the terms upper and lower bound were not used. The order of the questions was counterbalanced, resulting in six different orders. To determine whether participants thought the values within the interval were normally distributed, they were also asked to estimate the probability that the observed temperature would fall within four equal ranges between upper and lower bound values for the Saturday daytime high (see Figure 2(D)–(G)). Appendix A shows a full list of the questions asked.

Finally, participants answered questions about their age, coursework, and temperature scale preference (Fahrenheit, Celsius, or no preference). Between one and two questions were displayed on the screen at the same time. Participants were allowed to go back and change their answers to any previous questions at any point in the experiment.

2.1.4. Forecast Values. All forecast values were realistic and based on historic records for the region. For each task, the decision threshold value was within the interval in one forecast and outside of it in the other. Notice in Table 1 that the threshold for a freeze warning, 32°F (0°C), was within the interval for Friday night but below the interval for Saturday night. The forecast values for the hard freeze warning task were uncommonly cold winter temperatures for Washington State where the experiment was conducted. The threshold for a hard freeze warning, 25°F (−3.9°C), was within the interval for Friday night but below the interval for Saturday night. The forecast values for the heat-health-related task were

FIGURE 2: Ranges for which participants were asked to estimate the likelihood. Font sizes are slightly smaller in this figure than they were in the experiment to conform to journal requirements.

uncommonly warm summer temperatures for the region. The threshold for a heat warning, 100°F (37.8°C), was within the interval for Friday day but above the interval for Saturday day.

2.1.5. Forecast Format. The median value, shown alone in the point estimate condition, was represented by a number shown in black bold 12-point font located vertically in a box with higher position indicating higher temperature at a ratio of 10 pixels to 1 degree. The predictive interval also included a bracket whose end points were labeled with numbers, in 10-point font, indicating upper and lower bound temperatures of the 80% predictive interval (see Figure 1). One version was facing left (Figure 1(a)) and the other was facing right (Figure 1(b)). To the right of each forecast was a key that explained the median forecast, “The temperature will usually be closest to this value,” as well as the upper and lower bound,

"1 in 10 days like this, the temperature will be equal to or greater/less than this value." The point estimate condition did not have a key (Figure 1(c)).

2.1.6. Individual Differences Measures

Numeracy Test. The numeracy test had 11 items [14] with one question each about proportion, percent, and probability and eight questions about risk. See Appendix B for the full list of questions.

Working Memory Test. The ability to hold and process information on a conscious level, WMC, can be measured reliably using reading span assessments [33]. This task requires that participants make a semantic judgment while maintaining a memory load [34]. Here, participants read a set of unrelated sentences, comprising 10–15 words, and judged whether each one made sense. Half of the sentences made sense and half did not. After each semantic judgment, a letter appeared on the screen for 800 ms that participants were required to memorize. From three to seven sentence-letter pairs were presented in a set. After the final sentence-letter pair in each set, participants selected the letters remembered, from the entire set, in the order in which they were presented. The total number of letters recalled in the correct order, out of the 75 that were presented, constituted the score. For example, if a participant was given the letters S-N-L-P-R and recalled S-N-R-L-P, only the first two letters would be considered correct.

2.1.7. Design. The experimental portion of this study, involving the temperature decision tasks, employed a 3×6 between-groups incomplete factorial design with 18 conditions. Forecast format had three levels: (1) left facing bracket, (2) right facing bracket, and (3) a single-value point estimate (see Figures 1(a)–1(c)). The other independent variable was percent chance question order with six levels. All permutations of the three questions asking participants to indicate the percent chance that the observed temperature would be between the two bounds, above the upper bound, and below the lower bound were tested. The correlational portion examined the relationship between numeracy, working memory scores, and the rate of the two predictive interval interpretation errors described above.

2.2. Results. We conducted a series of analyses to determine whether predictive intervals improved decisions as compared to conventional point estimates. Then, we examined likelihood estimates to determine participants' impression of the distribution of possible outcomes. Finally, we examined interpretation errors to determine whether they were related to presentation format, numeracy, or WMC scores. All analyses excluded participants who failed to complete the tasks (11/644, 2%) or did not meet an 85% correct criterion on the semantic judgment (111/644, 17%) in the reading span WMC task, leaving a total of 526 participants. Similar proportions were below 85% in previous research [35]. This threshold is usually imposed to prevent participants from ignoring the semantic judgment task in favor of rehearsing the letter series.

2.2.1. Binary Decisions. Those using predictive intervals tended to take precautionary action more often when adverse events were likely, and less often when adverse events were unlikely (Figure 3(a)). Binary decisions were analyzed separately for the freeze warning, the hard freeze warning, and the heat-health tasks because the pattern might differ by weather scenario or task structure. The results were similar in analyses conducted with and without those who committed a DCE so all participants were used in the analyses reported here. In all subsequent analyses, the forecasts will be referred to by the median value temperature (see Table 1).

In the freeze warning task, the predictive interval increased cautiousness as compared to the point estimate when the threshold value (32°F; 0°C) was within the interval (33°F; 0.6°C forecast). A logistic regression analysis (appropriate for categorical outcomes such as the errors that were the dependent variable here) revealed that the predictive interval increased the likelihood of issuing a freeze warning by 2.01 times, $\text{Exp}(B) = 2.01$, $p < .001$, $\chi^2(1, N = 526) = 15.86$, $p < .001$, compared to the point estimate. For the 36°F (2.2°C) forecast, less than 10% of participants in either condition posted a warning and the difference was not significant. There was a similar pattern of results for the hard freeze warning task. The predictive interval increased the likelihood of issuing a freeze warning by 1.57 times, $\text{Exp}(B) = 1.57$, $p < .02$, $\chi^2(1, N = 526) = 5.58$, $p < .02$, compared to the point estimate when the threshold value (25°F; -3.9°C) was within the interval (26°F; -3.3°C forecast). For the 29°F (-1.7°C) forecast, less than 20% of participants in either condition posted a warning and the difference was not significant. Taken together, these results suggest that predictive intervals increase precautionary action specifically in situations in which the threshold for action is within the interval rather than in general.

However, there was a different pattern of results for the heat-health task in which the active choice, taking the elderly relative for a walk, was not precautionary. For the 99°F (37.2°C) forecast, the majority of participants in both conditions decided against taking the elderly person out and the difference was not significant. For the 96°F (35.6°C) forecast in which the threshold (100°F; 37.8°C) was above the interval however, the predictive interval increased the likelihood of making the less cautious choice by 2.34 times, $\text{Exp}(B) = 2.34$, $p < .001$, $\chi^2(1, N = 526) = 20.30$, $p < .001$, compared to the point estimate. Thus, the predictive interval also counteracted overcautiousness when the threshold was outside of the interval (Figure 3(a)). Overall, the predictive interval allowed participants to better discriminate between situations for taking and avoiding precautionary action than did the point estimate.

Likelihood Estimations. Participants in the predictive interval condition thought there was a *smaller* chance that the observed temperature would fall beyond the upper and lower bounds than did those using the point estimate, but estimated a *greater* percent chance than was intended by the forecast. Question order had no significant impact on likelihood estimates in any of the subsequent analyses and will not be discussed further.

FIGURE 3: The percentage of participants who took precautionary action in point estimate and predictive interval conditions over forecasts in Experiment 1 (a) and Experiment 2 (b). The threshold temperature was inside the predictive interval for the 33°F (0.6°C), 26°F (−3.3°C), and 99°F (37.2°C) forecasts. Stars indicate statistically significant differences.

In order to determine the impact of the predictive interval on participants' understanding of the likelihood of future temperature values, they were asked to indicate the percent chance that the observed temperature would be (1) *above* the upper bound, (2) *below* the lower bound, and (3) *between* the two boundaries for one Friday and one Saturday forecast in each of the three tasks. We calculated the mean percent chance selected in these three segments and compared the predictive interval to point estimate conditions using independent groups *t*-tests (Bonferroni's correction of $\alpha = .017$). The mean percent chance below the lower bound in the predictive interval condition was significantly smaller ($M = 15.33\%$, $SD = 11.52\%$) than in the point estimate condition ($M = 26.77\%$, $SD = 13.95\%$), $t(377.84) = 9.82$, $p < .001$, Cohen's $d = 1.01$. The mean percent chance above the upper bound was also significantly smaller in the predictive interval condition ($M = 14.99\%$, $SD = 10.66\%$) than in the point estimate condition ($M = 25.35\%$, $SD = 14.72\%$), $t(341.89) = 8.73$, $p < .001$, Cohen's $d = .94$. The mean percent chance between the upper and lower bounds was significantly larger in the predictive interval ($M = 78.51\%$, $SD = 14.68\%$) than in the point estimate condition ($M = 74.94\%$, $SD = 16.19\%$), $t(341.28) = 2.56$, $p = .01$, Cohen's $d = .25$. Thus, the predictive interval reduced expectations that the observation would fall beyond the interval, essentially narrowing the range of possible values.

Interestingly, those using the point estimate were influenced by the task. When the task concerned nighttime low temperatures, participants in the point estimate condition indicated a larger mean percent chance that the observation would fall below the lower bound than above the upper bound. This was true for both the freeze warning (lower: $M = 28.30\%$, $SD = 16.38\%$ versus upper: $M = 22.11\%$, $SD = 15.42\%$), $t(205) = 5.29$, $p < .001$, $d = .39$, as well as the hard freeze warning forecasts (lower: $M = 26.70\%$, $SD = 15.80\%$ versus upper $M = 23.98\%$, $SD = 15.77\%$), $t(205) = 2.86$, $p = .005$, Cohen's $d = .17$. However, when the task concerned daytime high temperatures, the pattern reversed

and participants in the point estimate condition indicated a larger mean percent chance above the upper bound ($M = 29.95\%$, $SD = 18.84\%$) than below the lower bound (25.32% , $SD = 16.29\%$), $t(205) = 4.65$, $p < .001$, Cohen's $d = .26$. There were no significant differences in upper and lower bound estimates among participants in the predictive interval condition suggesting that explicit uncertainty information protected them from task biases. All comparisons used a Bonferroni's correction of $\alpha = .008$.

2.2.2. Temperature Distribution. In order to determine participants' impression of the distribution of values within the interval, we compared the likelihood estimates for the central and outer segments within the predictive interval (see Figure 2(D)–(G)). A normal distribution would be suggested by greater likelihood estimates for values near the center of the predictive interval than further away. Indeed, this was the case, both among participants who saw the predictive interval and among those who did not. A 2 (within groups: central/outer segments) $\times$ 2 (between groups: point estimate/predictive interval forecast) mixed model ANOVA conducted on mean likelihood estimates revealed that the estimate for the combined central ranges ($M = 81.39\%$, $SD = 39.45\%$) was significantly larger than for the combined outer ranges ($M = 60.89\%$, $SD = 37.19\%$), $F(1, 524) = 302.04$, $p < .001$, $\eta^2_p = .37$. Participants using the predictive interval provided smaller estimates overall (136.53% , $SD = 74.90\%$) than did those using the point estimate (151.21% , $SD = 60.33\%$), $F(1, 524) = 5.58$, $p = .01$, $\eta^2_p = .01$. In addition, the inner and outer ranges were more similar in the predictive interval (central range $M = 74.06\%$, $SD = 40.09\%$; outer range $M = 62.47\%$, $SD = 38.73\%$) than in the point estimate condition (central range $M = 92.77\%$, $SD = 35.64\%$; outer range $M = 58.45\%$, $SD = 34.61\%$), $F(1, 524) = 74.00$, $p < .001$, $\eta^2_p = .12$. Indeed, 46% of participants using predictive intervals indicated that the percent chance for the central range was *equal* to the outer range compared to only 16%

FIGURE 4: Mean percent chance estimates by participants for values within the predictive interval in Experiment 1.

of point estimate condition participants, $t(516.14) = 7.99$, $p = .001$, $d = .70$ (see Figures 4(a) and 4(b)), suggesting that the predictive interval tends to equalize the likelihood of values within its boundaries.

Thus, predictive intervals benefit users by narrowing the range of expected outcomes allowing people to be more decisive than when this information is omitted. However, predictive intervals may have the unintended effect of implying a similar likelihood for outcomes within the interval. We will return to this issue in the Discussion.

2.2.3. Interpretation Errors

Deterministic Construal Errors. Next, we examined DCEs, misconstruing the upper bound value as the most likely daytime high (e.g., in Figure 1(a), selecting 44°F (6.7°C) as the most likely daytime high for Friday) or the lower bound value as the most likely nighttime low. The total number of DCEs, over the twelve relevant questions, was calculated for each participant and constitutes the dependent variable for the following analyses. Participants who used the predictive interval forecast made significantly more DCEs ($M = 2.35$, $SD = 3.55$) than participants who used the point estimate ($M = .23$, $SD = .59$), $t(346.29) = 10.48$, $p < .001$, Cohen's $d = .83$, suggesting that it was not a random error. Moreover, participants with DCEs made significantly different decisions than did those who correctly interpreted the predictive interval, suggesting that they believed the lower or upper bounds were the most likely temperatures. For two of the night time low temperature forecasts, those who committed DCEs were significantly more likely to issue a freeze warning than were those who correctly interpreted the predictive interval (33°F [0.6°C]: $\text{Exp}(B) = 6.88$, $p < .001$, $\chi^2(1, N = 526) = 46.97$, $p < .001$ and 26°F [-3.3°C]: $\text{Exp}(B) = 8.83$, $p < .001$, $\chi^2(1, N = 526) = 22.63$, $p < .001$). For the 99°F forecast, those who committed a DCE were significantly more likely to decide against the walk, $\text{Exp}(B) = 2.67$, $p < .05$, $\chi^2(1, N = 526) = 4.81$, $p < .03$.

We tested three possible explanations for DCEs: (1) the orientation of the bracket in the graphic, (2) limitation in numeracy skills, and (3) limitation in working memory

capacity. Because they address these explanations, subsequent analyses in this section will be confined to the predictive interval condition. The mean number of DCEs was comparable in the left ($M = 2.40$, $SD = 3.50$) and the right ($M = 2.30$, $SD = 3.62$) facing bracket conditions, $t(318) = 0.25$, $p = .81$, Cohen's $d = .03$, suggesting that DCEs were not due to reading order. For the remainder of the analyses, these two conditions were combined.

In order to determine whether DCEs were due to cognitive limitations, a binomial generalized linear model was conducted on the total number of errors, using numeracy and WMC as simultaneous predictor variables. Numeracy and WMC both significantly predicted the odds of correct interpretation (not making a DCE). For each additional unit increase in numeracy score, the likelihood of successful interpretation increased by 1.15 times, $\text{Exp}(B) = 1.15$, $p < .001$, $\chi^2(1, N = 320) = 30.65$, $p < .001$. For each additional unit increase in WMC score, the likelihood of successful interpretation increased by 1.01, $\text{Exp}(B) = 1.01$, $p < .001$, $\chi^2(1, N = 320) = 7.03$, $p < .001$. However, both effect sizes [36] were small ($d < .10$), suggesting that numeracy and working memory limitations provide only a partial explanation for these errors.

Subadditivity. Here, we define subadditivity as the sum of percent chance estimates (below, above, and between the boundaries) that exceeds 100%, because these three comprise all possible outcomes. A slight majority of participants ($N = 317/526$, 60%) made this error, although it did not significantly impact decisions. This was tested in a series of logistic regression analyses using participants estimate of the chance that observations would fall beyond the interval to predict their binary decision. Approximately a third of participants, 156/526 (30%), gave answers that summed to exactly 100. There were also a few participants whose answers summed to less than 100 and were removed ($N = 53/526$, 10%) from subsequent analyses because these errors were theoretically different so they could not be combined with subadditivity errors. However, they were too infrequent to analyze separately.

Subadditivity was reduced by the predictive interval format but not by higher numeracy or higher WMC. We

FIGURE 5: The mean percent chance estimate for events within the boundary values of the predictive interval by point estimate and predictive interval conditions plotted over whether a single question (packed) or four separate questions (unpacked) were asked in Experiment 1.

conducted a 2 (extreme quartile split on numeracy scores: high/low) $\times$ 2 (extreme quartile split on WMC scores: high/low) $\times$ 2 (forecast format: predictive interval/point estimate) ANOVA on percent chance estimate sums. Participants using the predictive interval had significantly smaller sums ($M = 113.66\%$, $SD = 27.52\%$) than did those using the point estimate ($M = 128.72\%$, $SD = 32.49\%$), $F(1, 201) = 7.69$, $p < .01$, $\eta^2_p = .04$, a medium effect size. Neither the main effect of numeracy ($F(1, 201) = 2.63$, $p = .11$, $\eta^2_p = .01$) nor working memory capacity ($F(1, 201) = .18$, $p = .67$, $\eta^2_p = .001$) reached significance. Nor were there any significant interactions.

However, those using the predictive interval may have overcome subadditivity simply because they were reading the correct values off the key, without necessarily understanding the principles involved. In order to test whether the advantage for the predictive interval in combatting subadditivity extended to values that were not explicitly mentioned in the key, we examined participants' likelihood estimates for temperatures occurring within the upper and lower bound values. This was asked as a single question (within boundaries) to which we refer as "packed." We also asked four separate questions (designed to answer the "normal distribution" question discussed above) essentially "unpacking" the interval. A sum was calculated over participants' answers to the four separate questions. Then, a mixed model ANOVA was conducted on the likelihood estimates. The within-groups variable was "packing" (single question: packed; four questions: unpacked). The between-groups factor was forecast format (predictive interval, point estimate). As Figure 5 shows, the unpacked sum ($M = 145.52\%$, $SD = 71.15\%$) was significantly greater than the packed sum ($M = 79.28\%$, $SD = 13.40\%$), $F(1, 471) = 421.18$, $p < .001$, $\eta^2_p = .47$. Moreover, there was a significant interaction between forecast format and packing, $F(1, 471) = 9.10$, $p < .003$, $\eta^2_p = .02$. In the predictive interval condition, the unpacked sum ($M = 139.19\%$, $SD = 76.55\%$) was more similar to the packed sum ($M = 80.94\%$, $SD = 12.18\%$) than in the point estimate condition

(unpacked: $M = 155.11\%$, $SD = 61.04\%$; packed: $M = 76.77\%$, $SD = 14.73\%$). This suggests that the predictive interval helped some people to overcome the subadditivity error by alerting them to the logical rules that apply in this situation.

2.3. Discussion. Experiment 1 confirmed a clear advantage for predictive interval forecasts. These results suggest that the predictive interval allows people to better identify situations in which to take precautionary action than does a point forecast, replicating previous research [12]. We refer to this advantage as an increase in decisiveness because those using the point estimate performed more similarly across forecasts whereas those using predictive intervals made decisions that differentiated to a greater degree between forecasts. Although this task had no economically rational threshold above which precautionary action is required, in other research in which such a threshold is calculable [37], the pattern was the same. People who had numeric uncertainty estimates performed more differently above and below the threshold than did those with point estimates. In addition, they made decisions that were closer to the economically rational standard.

The results reported here also suggest that the predictive interval increases decisiveness because it informs users of the boundaries beyond which outcomes are particularly unlikely. Participants using the predictive interval made likelihood estimates for values beyond the interval that were significantly smaller than the estimates provided by those using the point estimate. However, their estimates were larger than the intervals presented in the forecast, perhaps because they thought that forecasters were overconfident in their precision, a well established psychological bias [38] of which users may be intuitively aware. Notice also that the probabilities estimated by those using only the point estimate suggest, as has been shown in other research [39, 40], that people understand that forecasts include uncertainty even when it is not provided.

Moreover, Experiment 1 revealed that most participants understood that there was a greater chance that the observed temperature would be closer to the "most likely" value. However, this effect was reduced among those using the predictive interval, suggesting that, for at least some users, the predictive interval makes the likelihood of values within its boundaries seem similar.

One of the major goals of Experiment 1 was to better understand two misinterpretations that had been noted in previous research and were observed here as well. Of the two errors, only DCEs impacted user decisions, suggesting that they are more serious from a practical perspective. Indeed, DCEs were modestly related to individual differences. However, the explanatory power of these individual difference variables was limited, suggesting that interpretation errors were largely due to other factors. The likelihood estimation errors, sums greater than 100%, were not related to the individual difference measures nor were they an artifact of question order. Although such errors may betray a fundamental misunderstanding of probability, they did not appear to effect decisions, suggesting that these errors may not have a practical impact.

Interestingly, the predictive interval forecast itself appeared to promote a logical understanding of the forecast. The sums across 100% of the outcomes made by those using the predictive interval were smaller and closer to the correct answer than were the sums of those using the point estimate. Importantly, this was true of ranges of values not mentioned in the key, suggesting that the advantage was conceptual. Moreover, the predictive interval appeared to counteract task bias effects. While those in the point estimate condition estimated greater chances for lower temperatures in the freeze warning task and higher temperatures in the health-related decision, perhaps out of an abundance of cautiousness [41], estimates in the predictive interval condition were unbiased. Taken together, these results suggest that, rather than being confusing, the predictive interval tended to prompt more careful reasoning and provided participants with a better understanding of the uncertainty in the situation, at least among this college student sample.

Thus, Experiment 1 revealed a number of clear benefits to users of predictive interval forecasts. However, because the participants in this study were college students, they may have been better equipped to take advantage of predictive intervals than would the general public. A sample with a broader range of educational backgrounds and ages is required to determine whether these benefits are likely to extend to general public end-users. This was the primary goal of Experiment 2 described below.

3. Experiment 2

Experiment 2 was a web-based questionnaire using the same three decision tasks described in Experiment 1 to compare predictive intervals to point estimates in a nonstratified general population sample. The main question for Experiment 2 was whether this sample would be able to understand predictive intervals and to benefit from them in the same way as college students had done.

Of particular concern were the errors that had been observed among college students, which may increase among less educated users. In Experiment 2, we sought to reduce DCEs by improving the manner in which they were presented. Two additional presentation formats were tested. One was a graphic that directly contradicted the diurnal fluctuation misinterpretation (continuous range; see Figure 6(a)), by including a line connecting the daytime high and nighttime low. Perhaps when viewers realized that diurnal fluctuation was represented by the line, they would not attribute the same meaning to the predictive interval itself. In other words, this graphic may serve to block DCEs.

The second potential solution for DCEs is a predictive interval forecast expressed in an entirely textual format (see Figure 6(b)) shown to be effective in other situations [12]. Visualizations may promote DCEs because people tend to interpret the image as diurnal fluctuation before they have a chance to read the key. Once the initial misinterpretation is established, some users may be convinced that they understand the forecast and fail to make an effort to fully grasp the explanation in the key. Notice that because we seek to simulate a natural situation in which users would encounter

TABLE 2: Demographic information for participants in Experiment 2 compared to 2012 US Census.

	M-Turk sample	National population (2012 census data)
Age	Mean 35.11, median 32 years	Median 37.2 years
Gender	56.4% females, $n = 219$	50.8% females
Education	64.4% college educated, $n = 250$	39% college educated

this information in a web format, we do not require them to read the key specifically but rather to “use the forecast” which includes a key, to inform their decisions. Thus, the solution may be to eliminate the visualization altogether. Alternatively, the visualization may benefit the noncollege users among the sample tested here [42].

In addition, Experiment 2 reexamined subadditivity, likelihoods that summed to greater than 100%. Although this error did not affect decision quality among college students, it may be more prevalent and problematic in a general population. Experiment 2 tested whether subadditivity was affected by the definition provided in the key. Because frequency makes the reference class more prominent than does probability, it may reduce overestimation. Frequency may also be more understandable in general, reducing DCEs, especially for less educated users. Both frequency and probability were tested in Experiment 2 in order to determine which is better understood by a more diverse sample.

3.1. Method

3.1.1. Participants. Participants were recruited from Amazon’s Mechanical Turk (<https://www.mturk.com/mturk/welcome>), an online crowdsourcing service that hires workers for the execution of tasks (called Human Intelligence Task, or HIT). Participants were compensated \$0.10 for their responses. Only the 388 participants who were residents in the US had a 90% prior “approval rate” (percent of prior HITs accepted by requester) and had no professional atmospheric science background were included. Seventy-two percent of participants classified themselves as everyday weather forecast users, 21% as interested amateurs, and 7% as having some education involving atmospheric science. The vast majority, 91%, indicated that they preferred the Fahrenheit temperature scale. The sample was similar to the 2012 US Census [43] although there were slightly more females (56%), the average age was slightly younger (35 years), and the sample was better educated (see Table 2). Thirty-six percent of participants earned a high school diploma, 47% had a college degree, and 16% had an advanced degree. Only 1% had not completed high school compared to almost 12% nationally.

Procedure. The three weather-related decision tasks described in Experiment 1 (freeze warning, hard freeze warning, and heat health) were posted in an online questionnaire over a 10-week period between the months of August and November. Before participants linked to the externally hosted decision

(a) Continuous range

Forecast for Seattle, WA			
Fri Nov 30	Fri Nov 30	Sat Dec 1	Sat Dec 1
Daytime high will usually be closest to 41°F. 1 in 10 days like this, the daytime high will be equal to or greater than 44°F. 1 in 10 days like this, the daytime high will be equal to or less than 38°F.	Nighttime low will usually be closest to 33°F. 1 in 10 days like this, the nighttime low will be equal to or greater than 36°F. 1 in 10 days like this, the nighttime low will be equal to or less than 30°F.	Daytime high will usually be closest to 39°F. 1 in 10 days like this, the daytime high will be equal to or greater than 44°F. 1 in 10 days like this, the daytime high will be equal to or less than 34°F.	Nighttime low will usually be closest to 36°F. 1 in 10 days like this, the daytime high will be equal to or greater than 39°F. 1 in 10 days like this, the nighttime low will be equal to or less than 33°F.

(b) Text

(c) Point estimate

(d) Bracket

FIGURE 6: Visualizations used in the temperature decision tasks in Experiment 2. Font sizes are slightly smaller in this figure than they were in the experiment to conform to journal requirements.

tasks, they read an explanation of informed electronic consent. Each of the three tasks was accompanied by a two-day temperature forecast showing the daytime high and nighttime low temperature for consecutive days, with values identical to Experiment 1 (see Table 1). Participants indicated

the temperature they expected to observe for the decision-critical time periods, made a decision for each date, and answered the same set of questions about each forecast as they did in Experiment 1. The only difference was that the questions about the chance of observed temperatures in

FIGURE 7: Keys used to explain the predictive interval in Experiment 2. Font sizes are slightly smaller in this figure than they were in the experiment to conform to journal requirements.

ranges within the interval, asked about a single forecast in Experiment 1, were asked about three forecasts in Experiment 2 to ensure that the results were not forecast specific.

3.1.2. Forecast Formats. There were four forecast formats; two were identical to formats tested in Experiment 1, a single-value point estimate (Figure 6(c)) and the left facing predictive interval bracket (Figure 6(d)). Two additional predictive interval formats were tested as well. The first was a simple text-only expression in which the forecast values were shown below their definitions (Figure 6(b)). The median forecast was shown on the top and the upper and lower bound values were shown below it in that order. A new visualization, the “continuous range” format, was also tested. The daytime high and nighttime low temperatures were connected by lines that crossed the panels to illustrate diurnal fluctuation (see Figure 6(a)). The median forecasts were connected with a solid black line and the upper and lower bounds were connected with dashed grey lines.

All three predictive interval forecast formats were accompanied by a key defining each value. There were two key conditions (Figure 7), one with a frequency explanation (“1 in 10 days like this, the temperature will be equal to or greater (less) than this value”) and one with a probability explanation (“There is a 10% chance that the daytime high will be equal to or greater (less) than this value”). The median forecast was accompanied by this explanation: “The temperature will usually be closest to this value.”

3.1.3. Design. Participants were randomly assigned to one of six predictive interval forecast conditions (3×2 between-groups factorial design) or to a point estimate condition, which acted as a control. The independent variables in the predictive interval forecast condition were format with three levels (text, bracket, and continuous range) and key expressions with two levels (probability, frequency). The point estimate condition was presented without a key.

3.2. Results. We conducted a series of analyses to determine whether the advantages for predictive interval forecasts, in terms of both decision-making and uncertainty understanding, were replicated in this more diverse sample. Then, we examined the two interpretation errors testing whether any of the predictive interval formats or key explanations reduced errors. For all analyses, we also tested the effect of education as a dichotomized variable, those with a high school degree or less ($N = 138$) versus a college degree or more ($N = 250$). Because some education levels failed to meet minimum cell size, a more finely grained analysis was not possible.

3.2.1. Advantages of Predictive Intervals. All of the advantages for predictive interval forecasts, shown in Experiment 1, were replicated here. Moreover, the statistics are remarkably similar to Experiment 1. Education level did not have a significant impact.

3.2.2. Binary Decisions. Those using the predictive interval made better decisions overall, taking precautionary action more often when adverse events were likely and less often when adverse events were unlikely (Figure 3(b)). No significant differences due to predictive interval format or key explanation were predicted or observed with regard to binary decisions, so these variables will not be mentioned in the following analyses. Instead, predictive interval formats will be combined and compared to the point estimate control condition.

In the freeze warning task, for the 33°F (0.6°C) forecast, a logistic regression analysis revealed that the predictive interval increased the likelihood of issuing a freeze warning by 1.95 times, $\text{Exp}(B) = 1.95$, $p = .03$, compared to the point estimate. The effect of education was not significant ($\text{Exp}(B) = 1.08$, $p = .71$). For the 36°F (2.2°C) forecast, very few participants posted a warning and neither the difference in format, $\text{Exp}(B) = 1.61$, $p = .45$, nor education was significant, $\text{Exp}(B) = 1.03$, $p = .94$.

FIGURE 8: Mean percent chance estimates for values within the predictive interval (a) and point estimate (b) conditions in Experiment 2.

Similarly, in the hard freeze warning task, for the 26°F (−3.3°C) forecast, a logistic regression analysis revealed that the predictive interval forecast increased the likelihood of issuing a freeze warning by 2.04 times, $\text{Exp}(B) = 2.04$, $p = .02$, but the effect of education failed to reach significance, $\text{Exp}(B) = 1.52$, $p = .08$. For the 29°F (−1.7°C) forecast, less than 20% of participants in either format condition posted a warning. Neither the difference between formats, $\text{Exp}(B) = 1.18$, $p = .68$, nor education was significant, $\text{Exp}(B) = .89$, $p = .68$. Taken together, these results suggest that as with Experiment 1, predictive intervals increase precautionary action specifically in situations in which the threshold for action is within the interval rather than in general.

As with Experiment 1, there was a different pattern for the heat-health task, in which the active choice, taking the elderly relative for a walk, was not precautionary. For the 99°F (37.2°C) forecast, more than 70% of participants in both format conditions decided against the walk and the difference was not significant, $\text{Exp}(B) = 1.28$, $p = .48$, nor was the difference between education levels significant, $\text{Exp}(B) = 1.06$, $p = .81$. For the 96°F (35.6°C) forecast in which the threshold (100°F, 37.8°C) was above the interval, only 24% of participants using the predictive interval decided against the walk as compared to 42% of those using the point estimate. A logistic regression analysis revealed that the predictive interval forecast increased the likelihood of making the less cautious choice by 2.09 times, $\text{Exp}(B) = 2.08$, $p = .02$. The effect of education was not significant, $\text{Exp}(B) = 1.29$, $p = .29$. Thus, the predictive interval also counteracted overcautiousness when the threshold was outside of the interval.

3.2.3. Likelihood Estimates. As with Experiment 1, participants in the predictive interval condition thought there was a *smaller* chance that the observed temperature would fall beyond the upper and lower bounds than did those using the point estimate, but estimated a *greater* percent chance than was intended by the forecast (10%).

The average percent chance estimates, below, above, and between the interval boundary values in the combined predictive interval conditions, were compared to those in the point estimate condition with independent t -tests, using a Bonferroni's correction of $\alpha = .017$. The mean percent chance below the lower bound value was significantly smaller in the

predictive interval condition ($M = 17.18\%$, $SD = 16.14\%$) than in the point estimate condition ($M = 25.14\%$, $SD = 18.24\%$), $t(67.10) = 3.03$, $p = .003$, Cohen's $d = .74$. The mean percent chance above the upper bound was significantly smaller in the predictive interval condition ($M = 16.35\%$, $SD = 15.00\%$) than in the point estimate condition ($M = 23.58\%$, $SD = 15.16\%$), $t(386) = 3.28$, $p = .001$, Cohen's $d = .33$. There was no significant difference between these two groups in the mean percent chance between the boundaries (predictive interval, $M = 73.05\%$, $SD = 19.70\%$; point estimate, $M = 76.39\%$, $SD = 17.12\%$).

As with Experiment 1, estimates of those using the point estimate were influenced by the task. In the hard freeze warning task, participants using the point estimate estimated a larger mean percent chance below the lower bound ($M = 26.76\%$, $SD = 22.43\%$) than above the upper bound ($M = 18.89\%$, $SD = 16.01\%$), $t(53) = 3.21$, $p = .002$, Cohen's $d = .40$. There were no significant differences for participants in the predictive interval condition, suggesting that it protected them from task biases. All comparisons used a Bonferroni's correction of $\alpha = .008$.

3.2.4. Temperature Distribution. As with Experiment 1, participants regarded the distribution of values within the interval as roughly normal (see Figure 8). We calculated the mean percent selected for the inner and outer ranges within the interval, across questions, and compared them in a repeated measures ANOVA with forecast format (predictive interval or point estimate) and level of education (high school or college degree) as between-groups factors. Participants gave larger likelihood estimates for the combined central ranges (64.43%, $SD = 37.68\%$) than for the combined outer ranges (48.49%, $SD = 33.71\%$), $F(1, 384) = 147.95$, $p < .001$, $\eta_p^2 = .28$, suggesting a normal distribution. As with Experiment 1, those using the predictive interval provided smaller estimates overall (107.56%, $SD = 64.49\%$) than did those using the point estimate (146.11%, $SD = 69.65\%$), $F(1, 384) = 14.84$, $p < .001$, $\eta_p^2 = .04$. In addition, the two ranges were more similar (central range: $M = 60.43\%$, $SD = 35.96\%$; outer range: $M = 47.13\%$, $SD = 33.09\%$) in the predictive interval condition than in the point estimate condition (central range: $M = 89.17\%$, $SD = 38.96\%$; outer range: $M = 56.94\%$, $SD = 36.52\%$),

FIGURE 9: The mean percent chance estimate by participants for events within the boundary values of the predictive interval by point estimate and predictive interval conditions plotted over whether a single question (packed) or four separate questions (unpacked) were asked in Experiment 2.

$F(1, 386) = 24.46, p < .001, \eta^2_p = .06$ (see Figures 8 and 9). Indeed, 34% of those with predictive intervals indicated that the percent chance for the central range was equal to the outer range as compared to only 15% of point estimate condition participants, $t(86.36) = 3.55, p = .001$, Cohen's $d = .76$. Level of education did not affect likelihood estimates for central and outer ranges ($F(1, 384) = .47, p = .49, \eta^2_p = .001$), nor was there an interaction between forecast format and level of education ($F(1, 384) = .02, p = .88, \eta^2_p < .001$). There were no significant differences due to predictive interval visualization formats ($F(2, 261) = .31, p = .74, \eta^2_p = .002$) or between frequency and probability explanations ($F(1, 261) = .04, p = .85, \eta^2_p < .001$).

3.2.5. Interpretation Errors

Deterministic Construal Errors. Next, we examined total DCEs, misconstruing the upper bound as the most likely daytime high or the lower bound as the most likely nighttime low. An ANOVA was conducted on DCE totals with forecast format (predictive interval or point estimate) and level of education (high school or college degree) as between-groups factors. Mean total DCEs in the combined predictive interval conditions were significantly greater ($M = 1.04, SD = 2.22$) than in point estimate condition ($M = 0.19, SD = .59$), $F(1, 384) = 17.05, p < .01, \eta^2_p = .02$, suggesting that it was not a random error. Surprisingly, there were no significant differences in DCEs due to education ($F(1, 384) = .06, p = .81, \eta^2_p < .001$), nor was there a significant interaction between forecast format and education ($F(1, 384) = .78, p = .38, \eta^2_p = .002$). As with Experiment 1, those who made DCEs made significantly different decisions, suggesting that they believed the lower or upper bounds were the most likely temperatures, compared with those who interpreted the predictive interval correctly. For three of the four nighttime low temperature forecasts, those committing

the DCE were more likely to issue a warning than were those who interpreted the forecast correctly (33°F [0.6°C]: $\text{Exp}(B) = 18.22, p < .001, \chi^2(1, N = 329) = 46.59, p < .001$; 36°F [2.2°C]: $\text{Exp}(B) = 2.76, p = .03, \chi^2(1, N = 329) = 3.99, p < .05$; 26°F [−3.3°C]: $\text{Exp}(B) = 5.31, p = .03, \chi^2(1, N = 329) = 22.63, p < .001$). For the 96°F [35.6°C] daytime high forecast, those who committed a DCE were 2.73 times more likely than those who correctly interpreted the forecast, to decide against the walk, $\text{Exp}(B) = 2.73, p = .02, \chi^2(1, N = 329) = 5.24, p = .02$.

In order to determine whether any predictive interval expressions reduced DCEs, a univariate ANOVA was conducted on mean total DCEs, excluding the point estimate condition, with two independent variables, predictive interval format (text, bracket, and continuous range), and key explanation (probability and frequency). There was a main effect for predictive interval format, $F(2, 328) = 5.54, p < .01, \eta^2_p = .03$. Tukey's post hoc analyses revealed that participants with the text format ($M = 0.49, SD = 1.04$) made significantly fewer DCEs than did those with the continuous range ($M = 1.42, SD = 2.84; p = .005$) or those with the bracket ($M = 1.20, SD = 2.31; p = .04$). Notice that the error is reduced by more than half. There were no significant differences due to key explanation ($F(1, 328) = 2.38, p = .12, \eta^2_p = .007$).

Subadditivity. Next, we examined subadditivity, the degree to which the sum of participants' percent chance estimates (below, above, and between the boundaries) exceeded 100%. Almost half of the participants ($N = 176/388, 45\%$) made likelihood estimates that summed to greater than 100%. Approximately a third of participants, 147/388 (38%), gave answers that summed to exactly 100. There were also a few participants whose answers summed to less than 100 and were removed ($N = 65/388, 17\%$) from subsequent analyses.

Unlike Experiment 1, however, those making this error were slightly more cautious in some situations than those who did not. Participants with higher subadditivity sums were more likely to take precautionary action in response to the forecast in each task in which the threshold value was outside of the interval (36°F [2.2°C]: $\text{Exp}(B) = 1.02, p < .001, \chi^2(1, N = 323) = 19.73, p < .001$; 29°F [−1.7°C]: $\text{Exp}(B) = 1.02, p < .001, \chi^2(1, N = 323) = 15.35, p < .001$; 96°F [35.6°C]: $\text{Exp}(B) = 1.01, p = .002, \chi^2(1, N = 323) = 9.51, p = .002$), suggesting that they actually thought the outcome was more likely.

Again, subadditivity was reduced by the predictive interval. However, subadditivity was unaffected by predictive interval format or key explanation, so these conditions will be combined (an ANOVA restricted to participants who saw the predictive interval indicated that neither uncertainty visualization ($F(2, 267) = .75, p = .47, \eta^2_p = .006$) nor key explanation ($F(1, 267) = .26, p = .61, \eta^2_p = .001$) affected subadditivity).

An ANOVA was conducted on subadditivity sums with forecast format (predictive interval, point estimate) and level of education (high school or college degree) as between-groups variables. Participants in the predictive interval condition had lower sums ($M = 114.27\%, SD = 27.83\%$) that were

closer to the correct value than did participants in the point estimate condition ($M = 127.47\%$, $SD = 25.80\%$), $F(1, 319) = 9.09$, $p < .01$, $\eta^2_p = .03$. There was no difference in subadditivity sums due to education ($F(1, 319) = .11$, $p = .74$, $\eta^2_p < .001$), nor was there a significant interaction between education and forecast format ($F(1, 319) = .00$, $p = .99$, $\eta^2_p < .001$).

In order to test whether the advantage for predictive intervals in combatting subadditivity extended to values that were not explicitly mentioned in the key, we examined participants' likelihood estimates for temperatures occurring between the upper and lower bound values. This was asked as a single question ("packed": Figure 2 range (C)) as well as in four separate questions ("unpacked": Figure 2 ranges (D)–(G)), for which a sum was calculated. A mixed model ANOVA was conducted on the percent chance estimates. The within-groups variable was "packing" (packed versus unpacked). The between-groups factor was forecast format (predictive interval, point estimate). As Figure 9 shows, there was a main effect of packing. The sum of the percent chance unpacked ($M = 119.58\%$, $SD = 69.11\%$) was significantly larger than that of the packed version ($M = 78.48\%$, $SD = 13.91\%$), $F(1, 321) = 87.73$, $p < .001$, $\eta^2_p = .22$. Moreover, there was a main effect of forecast format. Participants using the predictive interval format indicated significantly smaller values ($M = 96.49\%$, $SD = 40.55\%$) than those using the point estimate ($M = 112.89\%$, $SD = 43.73\%$), $F(1, 321) = 10.12$, $p = .002$, $\eta^2_p = .03$. There was a significant interaction between format and packing, $F(1, 321) = 8.93$, $p = .003$, $\eta^2_p = .03$. In the predictive interval condition, there was a smaller difference between the packed ($M = 78.51\%$, $SD = 13.54\%$) and unpacked ($M = 114.48\%$, $SD = 67.55\%$) conditions than in the point estimate condition (packed: $M = 78.33\%$, $SD = 15.90\%$; unpacked: $M = 147.45\%$, $SD = 71.55\%$). This again suggests that the predictive interval helps some people to overcome the subadditivity error.

3.3. Discussion. Experiment 2 confirmed a clear advantage for predictive interval forecasts among a diverse sample of nonexpert users. As with Experiment 1, participants were more decisive when using predictive intervals as compared to point estimates. As with Experiment 1, this appears to be because predictive intervals led users to expect a smaller chance of observations beyond the interval boundary values. Again, predictive intervals appeared to reduce subadditivity and task bias effects. Importantly, none of these advantages was affected by the level of education. This constitutes strong evidence that predictive interval forecasts can be useful to the general public in a range of situations. It is clear that none of these results was due to the new continuous range visualization or frequency explanation because there were no significant differences between these conditions and those also tested in Experiment 1 (bracket, probability explanation).

The two errors noted in previous research among college students were also found with this broader sample. Surprisingly however, the proportions of errors were comparable (or less) to that of college students and not related to education. In Experiment 2, subadditivity as well as DCEs influenced users' decisions, causing them to be more cautious in some

situations. Contrary to our hypothesis, neither error was reduced by the frequency explanation. Moreover, although we expected the continuous range to reduce DCEs because it highlighted the fact that diurnal fluctuation is indicated in the relationship between the daytime high and nighttime low temperatures, this was not the case. DCEs were only significantly reduced in the text condition, suggesting that visualizations in general encourage DCEs.

4. Conclusion

Because the predictive interval is a complex form of uncertainty information, it was an open question whether previous advantages found among college students would extend to a more diverse sample. Surprisingly, both the degree of benefit and the proportion of misinterpretation errors were comparable across experiments and none were influenced by level of education. In other words, those with less education (high school or less) were no more inclined to misinterpret the forecast and just as likely to benefit from it, as were the original college student participants. It is important to note however that only 1% of this M-Turk sample lacked a high school diploma as compared to almost 12% nationally. Thus, although these results are encouraging, it is not clear whether they extend to our least educated citizens.

Two additional benefits for predictive intervals were discovered in the experiments reported here. First, in both experiments, predictive interval forecasts protected users from task bias effects. Those using the point estimate tended to think it was more likely that observations would fall in the decision-critical range, below for the freeze warning tasks and above for the heat-health decision. Those using the predictive interval thought that outcomes in either direction were roughly equivalent regardless of the task. In addition, in both experiments, the predictive interval reduced subadditivity, suggesting that it served to alert some users to a logical analysis of the situation.

Most participants assumed that outcomes would be roughly normally distributed around the median value forecast. However, the inferred distribution in both experiments was flattened in the predictive interval condition and a substantial minority thought the likelihood was equal across the interval. This suggests that predictive intervals may have an unintended effect to standardize the likelihood of outcomes within its boundaries, an issue that should be more fully explored in future research.

The misinterpretations noted in previous research persisted here in roughly the same proportions. However, interpretation errors were only modestly related to WMC and numeracy. Because the proportion of errors was similar in the second experiment with a more representative sample and because of the absence of education effects in that experiment, we speculate that even though it was not possible to assess numeracy and WMC in the web-based study, the relationship would be similar. This suggests that limitations in WMC and numeracy are not major reasons that these particular errors occur. It may be that the interpretation errors noted in these experiments arise from common heuristics that reduce processing load in real-world decision

environments. It is important to note that in the majority of cases these errors were irrelevant to users' decisions or caused only a small increase in cautiousness. Thus, it may be a worthwhile trade-off to preserve cognitive resources for the tasks that matter most. In other words, people may not put forth the effort to overcome these errors because they do not have a critical impact on the decision at hand.

Another major concern was the misinterpretation of the predictive interval as a point estimate with additional information about diurnal fluctuations (DCE) because it served to obscure the uncertainty altogether. As with previous research, a substantial minority of participants exhibited this misunderstanding. DCEs were not reduced by reversing the bracket or by the continuous range visualization. This is clearly a persistent error in interpretation and may be related to a psychological "desire for certainty" [44, 45] and the more general tendency for "attribute substitution," in which people tend to automatically substitute an easier interpretation for a more difficult one [46]. Thus, people may have a natural, perhaps unconscious preference for the point estimate interpretation because of the reduction in processing load [47].

In addition, the results reported here suggest that text expressions block DCEs and all but eliminate the error. This result has important implications for uncertainty visualization in general, adding to a growing body of research suggesting that uncertainty visualizations may be problematic [12]. When a visualization is added, some people may notice it first and assume they know what it means (e.g., diurnal fluctuation) without fully processing the rest of the information. Combined with a "desire for certainty," this tendency could make the deterministic construal error particularly likely for many uncertainty visualizations. For that reason, we recommend that predictive intervals be communicated using a text format and call for additional research into uncertainty visualizations in general.

In sum, the predictive interval is a particularly adaptable form of uncertainty information because the forecaster needs not know the user's specific threshold of concern to provide relevant information about the likelihood of critical events. As such, the same forecast can serve a variety of users. The research reported here suggests that a broad range of general public end-users could understand predictive interval forecasts, without lengthy explanation or training. We do not claim that this is a sophisticated or theoretical understanding, but rather a "working" understanding that allows them to use predictive intervals to make better decisions tailored to their own circumstances and risk tolerance.

Appendix

A. Temperature Visualization Task Questions

Questions (1), (3), (5), (6), (7), and (11) correspond to the dependent variable temperature estimate, Questions (2) and (4) to binary decision, and Questions (8) through (10) and Questions (12) through (19) to understanding the uncertainty

- (1) What do you think the nighttime low temperature will be on Friday?

- (2) Do you want to issue a freeze warning (indicating temperature is expected to fall below 32°F) for the night of Friday?
- (3) What do you think the nighttime low temperature will be on Saturday?
- (4) Do you want to issue a freeze warning (indicating temperature may fall below 32°F) for Saturday night?
- (5) What do you think the daytime high temperature will be on Friday?
- (6) What do you think the daytime high temperature will be on Saturday?
- (7) What is the most certain daytime high temperature for Friday?
- (8) How certain is it that the daytime high temperature on Friday will be 44°F or higher?
- (9) How certain is it that the daytime high temperature on Friday will be between 38°F and 44°F?
- (10) How certain is it that the daytime high temperature on Friday will be around 38°F or lower?
- (11) What is the most certain nighttime low temperature for Saturday?
- (12) How certain is it that the daytime high on Saturday will be between 42°F and 44°F?
- (13) How certain is it that the daytime high on Saturday will be between 40°F and 42°F?
- (14) How certain is it that the daytime high on Saturday will be between 36°F and 38°F?
- (15) How certain is it that the daytime high on Saturday will be between 34°F and 36°F?
- (16) How certain is it that the nighttime low on Saturday will be 39°F or higher?
- (17) How certain is it that the nighttime low on Saturday will be between 33°F and 39°F?
- (18) How certain is it that the nighttime low on Saturday will be 33°F or less?
- (19) Which of the Saturday forecasts has the most uncertainty?

B. Numeracy Questions

- (1) Which of the following numbers represents the biggest risk of getting a disease? 1 in 100, 1 in 1000, 1 in 10
- (2) Which of the following represents the biggest risk of getting a disease? 1%, 10%, 5%
- (3) If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 100?
- (4) If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000?

- (5) If the chance of getting a disease is 20 out of 100, this would be the same as having a —% chance of getting the disease.
- (6) If Person A's risk of getting a disease is 1% in ten years, and Person B's risk is double that of A's, what is B's risk? 83
- (7) If Person A's chance of getting a disease is 1 in 100 in ten years, and Person B's risk is double that of A, what is B's risk?
- (8) In the Big Bucks Lottery, the chances of winning a \$10.00 prize are 1%. What is your best guess about how many people would win a \$10.00 prize if 1,000 people each buy a single ticket from Big Bucks?
- (9) Imagine that we roll a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?
- (10) The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?
- (11) In the Acme Publishing Sweepstakes, the chance of winning a car is 1 in 1,000. What percent of tickets of Acme Publishing Sweepstakes win a car?

Disclosure

This paper was based on the doctoral dissertation of Margaret A. Grounds.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] R. Gonzalez and G. Wu, "On the shape of the probability weighting function," *Cognitive Psychology*, vol. 38, no. 1, pp. 129–166, 1999.
- [2] J. J. Koehler, "When are people persuaded by DNA match statistics?" *Law and Human Behavior*, vol. 25, no. 5, pp. 493–513, 2001.
- [3] A. Tversky and D. Kahneman, "Judgment under uncertainty: heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.
- [4] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *American Association for the Advancement of Science. Science*, vol. 211, no. 4481, pp. 453–458, 1981.
- [5] A. Tversky and D. Kahneman, "Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment," *Psychological Review*, vol. 90, no. 4, pp. 293–315, 1983.
- [6] J. Baron, *Thinking and Deciding*, Cambridge University Press, New York, NY, USA, 3rd edition, 2000.
- [7] W. G. Cole, "Understanding Bayesian reasoning via graphical displays," in *Proceedings of the 1989 SIGCHI Conference on Human Factors in Computing Systems, CHI 1989*, pp. 381–386, usa, June 1989.
- [8] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decisions under risk," *Econometrica*, vol. 47, pp. 313–327, 1979.
- [9] R. A. LeBoeuf and E. Shafir, "The conflicting choices of alternating selves," in *Warrington College of Business*, University of Florida, Fla, USA, 2005.
- [10] K. E. Stanovich, *Who is rational? studies of individual differences in reasoning*, Lawrence Erlbaum Associates, Inc, Mahwah, NJ, USA, 1999.
- [11] S. Joslyn and J. LeClerc, "Decisions With Uncertainty: the glass half full," *Current Directions in Psychological Science*, vol. 22, no. 4, pp. 308–315, 2013.
- [12] S. Savelli and S. Joslyn, "The advantages of predictive interval forecasts for non-expert users and the impact of visualizations," *Applied Cognitive Psychology*, vol. 27, no. 4, pp. 527–541, 2013.
- [13] S. Joslyn, L. Nemec, and S. Savelli, "The benefits and challenges of predictive interval forecasts and verification graphics for end users," *Weather, Climate, and Society*, vol. 5, no. 2, pp. 133–147, 2013.
- [14] I. M. Lipkus, G. Samsa, and B. K. Rimer, "General performance on a numeracy scale among highly educated samples," *Medical Decision Making*, vol. 21, no. 1, pp. 37–44, 2001.
- [15] M. J. Kane, M. K. Bleckley, A. R. A. Conway, and R. W. Engle, "A controlled-attention view of working-memory capacity," *Journal of Experimental Psychology: General*, vol. 130, no. 2, pp. 169–183, 2001.
- [16] K. Broad, A. Leiserowitz, J. Weinkle, and M. Steketee, "Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season," *Bulletin of the American Meteorological Society*, vol. 88, no. 5, pp. 651–667, 2007.
- [17] A. H. Murphy, S. Lichtenstein, B. Fischhoff, and R. L. Winkler, "Misinterpretations of precipitation probability forecasts," *Bulletin of the American Meteorological Society*, vol. 61, pp. 695–1701, 1980.
- [18] S. L. Joslyn, L. Nadav-Greenberg, M. U. Taing, and R. M. Nichols, "The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision," *Applied Cognitive Psychology*, vol. 23, no. 1, pp. 55–72, 2009.
- [19] G. Gigerenzer, R. Hertwig, E. Van Den Broek, B. Fasolo, and K. V. Katsikopoulos, "'A 30% chance of rain tomorrow': How does the public understand probabilistic weather forecasts?" *Risk Analysis*, vol. 25, no. 3, pp. 623–629, 2005.
- [20] A. M. MacEachren, "Visualizing Uncertain Information," *Cartographic Perspectives*, no. 13, pp. 10–19, 1992.
- [21] E. R. Tufte, *Beautiful Evidence*, Graphics Press, Cheshire, UK, 2006.
- [22] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete, "Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2769–2778, 2012.
- [23] A. Tversky and D. J. Koehler, "Support theory: A nonextensional representation of subjective probability," *Psychological Review*, vol. 101, no. 4, pp. 547–567, 1994.
- [24] R. Hastie and R. M. Dawes, *Rational Choice in An Uncertain World: The Psychology of Judgment and Decision Making*, Sage Publications, Thousand Oaks, Calif, USA, 2001.
- [25] S. Sloman, E. Wisniewski, Y. Rottenstreich, C. Hadjichristidis, and C. R. Fox, "Typical versus atypical unpacking and superadditive probability judgment," *Journal of Experimental Psychology: Learning Memory and Cognition*, vol. 30, no. 3, pp. 573–582, 2004.

- [26] G. Gigerenzer and U. Hoffrage, "How to improve Bayesian reasoning without instruction: Frequency formats," *Psychological Review*, vol. 102, no. 4, pp. 684–704, 1995.
- [27] R. Hertwig and G. Gigerenzer, "The conjunction fallacy revisited: how intelligent inferences look like reasoning errors," *Journal of Behavioral Decision Making*, vol. 12, no. 4, pp. 275–305, 1999.
- [28] L. F. Rinne and M. M. M. Mazzocco, "Inferring uncertainty from interval estimates: effects of alpha level and numeracy," *Judgment and Decision Making*, vol. 8, no. 3, pp. 330–344, 2013.
- [29] S. Tak, A. Toet, and J. Van Erp, "The perception of visual uncertainty representation by non-experts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 6, pp. 935–943, 2014.
- [30] K. E. Stanovich and R. F. West, "Advancing the rationality debate," *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 701–717.
- [31] W. De Neys and N. Verschueren, "Working memory capacity and a notorious brain teasers: the case of the monty hall dilemma," *Experimental Psychology*, vol. 53, no. 2, pp. 123–131, 2006.
- [32] M. R. P. Dougherty and J. Hunter, "Probability judgment and subadditivity: the role of working memory capacity and constraining retrieval," *Memory and Cognition*, vol. 31, no. 6, pp. 968–982, 2003.
- [33] A. R. A. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: a methodological review and user's guide," *Psychonomic Bulletin & Review*, vol. 12, no. 5, pp. 769–786, 2005.
- [34] N. Unsworth, R. P. Heitz, J. C. Schrock, and R. W. Engle, "An automated version of the operation span task," *Behavior Research Methods*, vol. 37, no. 3, pp. 498–505, 2005.
- [35] R. P. Heitz and R. W. Engle, "Focusing the spotlight: individual differences in visual attention control," *Journal of Experimental Psychology: General*, vol. 136, no. 2, pp. 217–240, 2007.
- [36] J. Sánchez-Meca, F. Marín-Martínez, and S. Chacón-Moscoso, "Effect-size indices for dichotomized outcomes in meta-analysis," *Psychological Methods*, vol. 8, no. 4, pp. 448–467, 2003.
- [37] S. L. Joslyn and J. E. LeClerc, "Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error," *Journal of Experimental Psychology: Applied*, vol. 18, no. 1, pp. 126–140, 2012.
- [38] J. E. Russo and P. J. H. Schoemaker, "Managing overconfidence," *Sloan Management Review*, vol. 33, pp. 7–17, 1992.
- [39] S. Joslyn and S. Savelli, "Communicating forecast uncertainty: public perception of weather forecast uncertainty," *Meteorological Applications*, vol. 17, no. 2, pp. 180–195, 2010.
- [40] R. E. Morss, J. L. Demuth, and J. K. Lazo, "Communicating uncertainty in weather forecasts: a survey of the U.S. public," *Weather and Forecasting*, vol. 23, no. 5, pp. 974–991, 2008.
- [41] E. U. Weber, "From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events," *Psychological Bulletin*, vol. 115, no. 2, pp. 228–242, 1994.
- [42] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer, "Using icon arrays to communicate medical risks: overcoming low numeracy," *Health Psychology*, vol. 28, no. 2, pp. 210–216, 2009.
- [43] "US Census Bureau, Education attainment by selected characteristics: 2010," Retrieved from http://www.census.gov/compendia/statab/cats/education/educational_attainment.html.
- [44] S. Lichtenstein and P. Slovic, "Reversals of preference between bids and choices in gambling decisions," *Journal of Experimental Psychology*, vol. 89, no. 1, pp. 46–55, 1971.
- [45] S. Lichtenstein and P. Slovic, "Response induced reversals of preference in gambling: an extended replication in Las Vegas," *Journal of Experimental Psychology*, vol. 101, no. 1, pp. 16–20, 1973.
- [46] D. Kahneman and S. Frederick, "Representativeness revisited: attribute substitution in intuitive judgment," in *Heuristics and Biases*, T. Gilovich, D. Griffin, and D. Kahneman, Eds., pp. 49–81, Cambridge University Press, New York, NY, USA, 2002.
- [47] S. A. Sloman, "The empirical case for two systems of reasoning," *Psychological Bulletin*, vol. 119, no. 1, pp. 3–22, 1996.