The Benefits and Challenges of Predictive Interval Forecasts and Verification Graphics for End Users

SUSAN JOSLYN

University of Washington, Seattle, Washington

LOU NEMEC

United States Military Academy, West Point, New York

SONIA SAVELLI

University of Washington, Seattle, Washington

(Manuscript received 31 January 2012, in final form 29 November 2012)

ABSTRACT

Two behavioral experiments tested the use of predictive interval forecasts and verification graphics by nonexpert end users. Most participants were able to use a simple key to understand a predictive interval graphic, showing a bracket to indicate the upper and lower boundary values of the 80% predictive interval for temperature. In the context of a freeze warning task, the predictive interval forecast narrowed user expectations and alerted participants to the possibility of colder temperatures. As a result, participants using predictive intervals took precautionary action more often than did a control group using deterministic forecasts. Moreover, participants easily understood both deterministic and predictive interval verification graphics, based on simple keys, employing them to correctly identify better performing forecast periods. Importantly, participants with the predictive interval were more likely than those with the deterministic forecast to say they would use that forecast type in the future, demonstrating increased trust. Verification graphics also increased trust in both predictive interval and deterministic forecasts when the effects were isolated from familiarity in the second study. These results suggest that forecasts that include an uncertainty estimate might maintain user trust even when the single-value forecast fails to verify, an effect that may be enhanced by explicit verification data.

1. Introduction

Although weather forecasts continue to improve, user complaints continue as well, fueled by high-profile misses such as ''No-maggeden,'' the name given by *The Washington Post* to a major snowstorm predicted for the east coast of the United States during the Christmas holiday season of 2010. Despite the fact that the initial forecast was given a ''low confidence'' rating and it continued to be downgraded over subsequent days, in the end the focus was on the fact that the major storm that was initially forecasted failed to materialize. Many believe that mistrust caused by such misses could be reduced if forecasts included specific numeric uncertainty estimates (e.g., National Research Council 2006). Indeed, it is now clear that everyday users understand that all forecasts involve uncertainty, expecting a wide range of values even when given a deterministic forecast comprising a single-value such as a nighttime low temperature of 32°F (Joslyn and Savelli 2010). Acknowledging the uncertainty explicitly (e.g., 30% chance) might convey the notion that the forecast was *intended* as probabilistic and should be evaluated as such. Perhaps, then, forecasts that were previously regarded as ''misses'' would be seen as reliable.

In addition to increasing user trust, uncertainty forecasts provide potentially useful information about the likelihood of various possible outcomes that could inform forecast-related decision making. There is now

Corresponding author address: Susan Joslyn, Department of Psychology, University of Washington, P.O. Box 351525, Seattle, WA 98195.
E-mail: susanj@uw.edu

DOI: 10.1175/WCAS-D-12-00007.1

strong behavioral evidence that decisions based on forecasts including the probability of freezing are economically superior to decisions based on deterministic temperature forecasts when the decision task involves a freeze warning (Joslyn and LeClerc 2012; Roulston et al. 2006).

For situations in which the forecaster does not know the users' threshold of concern, uncertainty can be expressed as a range of values (Murphy and Winkler 1974), derived from forecast ensembles and referred to as a predictive interval (Raftery et al. 2005). A predictive interval provides the upper and lower boundaries of the range within which the observed value is expected with a specified probability, indicating, for instance, that there is an 80% chance that the nighttime low temperature will be between 30° and 36°F. Predictive intervals could be useful to decision makers with various parameter concerns and tolerances for risk. Furthermore, although predictive intervals are conceptually complex, there is preliminary evidence that nonexperts understand them when they are defined in simple terms compatible with the task at hand (Joslyn et al. 2009). It is not yet known, however, whether decisions based on predictive interval forecasts are significantly different than decisions based on conventional single-value forecasts. Nor is it known whether predictive interval forecasts inspire trust. These two issues were addressed in the work presented here.

It is possible that people will mistrust predictive interval forecasts to the same degree that they mistrust deterministic forecasts. They might think, for instance, that the forecast has described a range that is too narrow. In other words, given an 80% predictive interval, they may understand what is intended but expect the observed temperature to fall within the boundaries only 50% of the time. In fact, there is evidence that people often describe boundaries that are too narrow for confidence intervals. Confidence intervals are the minimum and maximum values thought to contain a particular number such as the population of Spain, with a certain probability. When people designate such values the correct answer tends to fall beyond the boundaries more often than is indicated by the probability, suggesting overconfidence (Alpert and Raiffa 1982), although there are other possible interpretations for this effect (e.g., Erev et al. 1994).

Thus, people may think that forecasters are overconfident when providing predictive intervals. If so, users may require verification showing the frequency of observed events relative to the forecasted probability over a large set of events to determine whether the forecast is ''well calibrated'' (i.e., the relative frequency matches the probability). As far as we know, however, the impact of verification on trust in either deterministic or probabilistic forecasts has not been tested empirically. This issue was also addressed in the work presented here.

To summarize, the work presented here had three major goals. The first was to test whether untrained users understand predictive interval forecasts and verification expressed here in simple graphics with definition keys. If so, this kind of information could be provided in a web format serving a wide range of users. The second goal was to determine whether predictive intervals influence user decisions. The third goal was to test the impact of predictive intervals and verification graphics on user trust. Both experiments involved a realistic freeze warning task embedded in an agricultural context. The first experiment provided an initial test of the forecasts and graphics. Based on the results, the definition provided in the key was refined and the procedure was simplified to conduct the second experiment.

2. Experiment 1

a. Method

1) Design

The experiment was a two by two full factorial design. Participants were randomly assigned to either a predictive interval or a deterministic forecast condition. Within each forecast format condition, participants were randomly assigned to either verification or no verification. The dependant variables were expected temperatures, freeze warning decision, uncertainty ratings, forecast performance evaluation, and a decision about using the forecast type again.

2) Participants

Participants were 302 University of Washington introductory psychology students of whom 54% were women. They ranged in age from 18 to 49 years and earned course credit for their participation.

3) Procedure

The experimenter administered informed consent procedures and read general instructions describing the goal of the computerized task (see the appendix). Participants were to decide whether to issue a freeze warning for an agricultural community so that farmers could protect their crops from temperatures at or below 32°F.[1] They were cautioned against posting the warning

[1] Water freezes at 32°F.

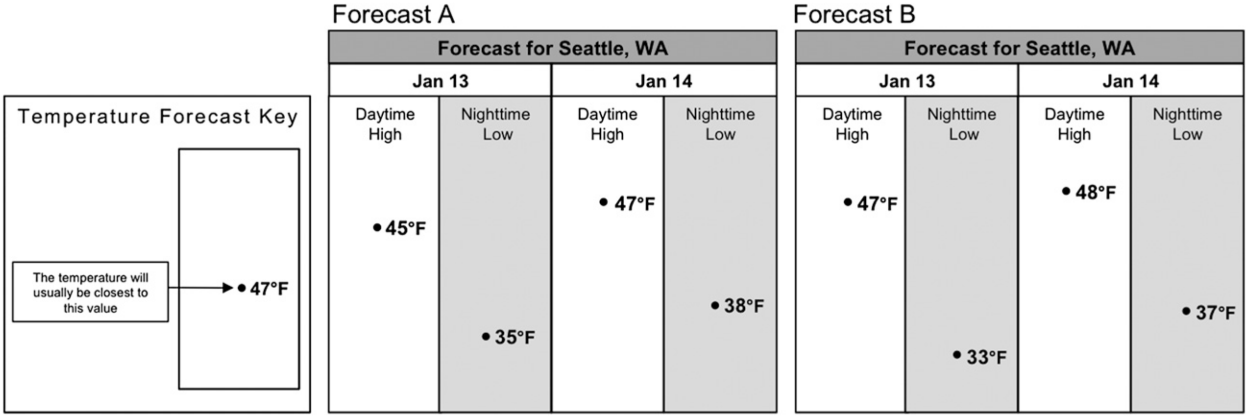

FIG. 1. Deterministic forecast graphics for experiment 1.

when freezing temperatures were not expected because crop protection involved material and labor costs.

After the instruction phase, participants saw a forecast graphic for 13 and 14 January that included daytime high and nighttime low temperatures for each date. The forecast was presented in one of two formats, a conventional single-value deterministic temperature forecast (Fig. 1) or a format that included the same single value as well as the temperatures at the upper and lower boundaries of the 80% predictive interval (Fig. 2). Participants were given no explicit instruction in decoding the graphics beyond the keys that accompanied them. Participants were to use the forecast to decide whether to issue a freeze warning for each night. Then, participants indicated the temperature they expected to observe for all four predictions (two daytime highs, two nighttime lows). For the nighttime low temperatures they also indicated the highest and lowest temperature that they would not be surprised to observe. The forecast graphic remained on the screen as participants worked through related questions. Questions targeting a particular time period (e.g., daytime high for 14 January) were shown simultaneously. Finally, to determine whether participants trusted the forecast, they were asked whether they would choose to use the same kind of forecast again. We selected this operationalization of trust for two reasons. The first is that a direct question about ''trust'' could have multiple interpretations. In addition, the primary concern from a practical perspective is the translation of trust into action: Will people actually use the information?

When all of these questions were answered, the forecast disappeared and a new set of questions was displayed. Participants in the verification condition saw a verification graphic and related questions while others, acting as a control group, saw a set of demographic questions, described below. The verification graphic showed the predicted and observed values for the previous 14 nights (1–14 January). There were two kinds of verification. Those in the deterministic-verification condition saw a comparison of the deterministic forecast to the observation, depicted visually as the distance

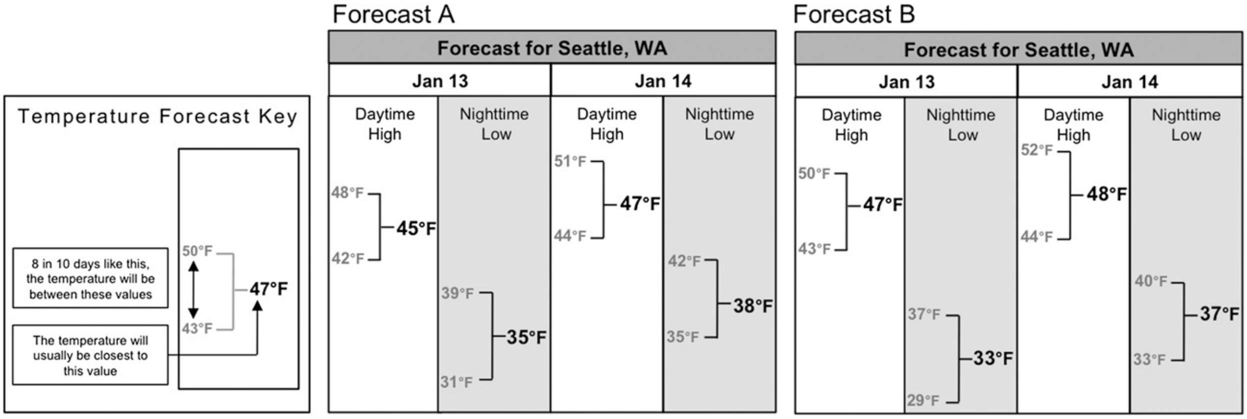

FIG. 2. Predictive interval forecast graphics for experiment 1.

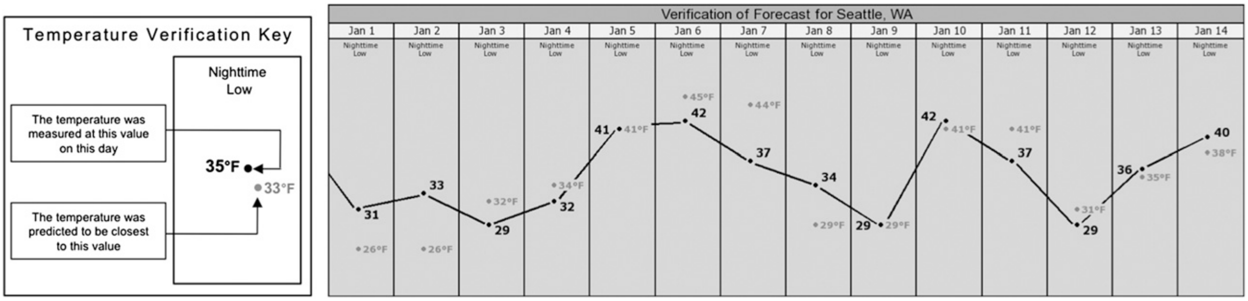

FIG. 3. Deterministic forecast verification graphic for experiment 1.

between the two and referred to as ''error'' (Fig. 3). Those in the predictive interval-verification condition saw a depiction of the ''calibration'' of the probabilistic forecast, depicted visually as the proportion of observations that fell within the predictive interval over a two-week period[2] (Fig. 4). All participants in the verification condition answered three questions requiring them to use the verification graphics to evaluate forecast performance. For the first question they rated performance over the entire two-week period on a scale of 1 (outstanding) to 7 (terrible). For the second they identified the single week that represented better forecast performance. The second week was both better calibrated and had the smallest error. For the third question participants ranked three individual days (1 January, 8 January, and 9 January). On one day, clearly the best by any standard (9 January in Figs. 3 and 4), the observed temperature and deterministic forecast were the same. For both of the other two days the error was four degrees, although on one day the observation fell within the predictive interval (1 January) and on the other day it was outside of the interval (8 January).

Next the screen was cleared again and a forecast for 28 and 29 January was shown. It was identical in format to the first forecast and participants answered the same questions as well as three additional questions about forecast uncertainty. Participants were asked to estimate the probability that the observed temperature would be at or above the temperature at the upper bound and at or below the temperature at the lower bound of the predictive interval. They were also asked to estimate the probability that the observed temperature would be between the two. Only temperature values were mentioned in these questions. The terms ''upper bound'' and ''lower bound'' were not used. Participants indicated their answers on a drop-down menu with choices from 0% to 100% in 10% increments. All of the uncertainty questions were also asked of participants in the deterministic condition to ascertain whether the predictive interval had an influence over pre-experimental expectations. Then, for those in the verification conditions, a second verification graphic (15–29 January) was presented and participants answered the same questions as for the first verification period.

To hold constant the time between the first and second forecasts, participants in the no-verification control condition answered questions about their age and educational background between the two forecasts. Participants in the verification conditions answered these questions last.

4) WEATHER DATA

The forecasts and observations were based on archived forecast data for the month of January 2009 from a local weather station. The month was split into two 14-day periods referred to here as weather datasets A and B. To test whether participants were sensitive to predictive interval reliability, each 14-day period was duplicated and adjusted so that one version of each had 79% (11/14) of the observations within the 80% predictive interval (well calibrated) and the other had 64% (9/14) of the observations within the 80% predictive interval (ill calibrated). In all other respects they were similar. The mean predictive interval width for all four periods was 8°F. The mean forecast error (predicted minus observed temperature) was similar for the 79% (SE = 3.77°F) and the 64% (SE = 3.78°F) calibration levels. For all four periods the average error above the single-value forecast was equal to the average error below it. All four periods had two days on which the observed temperature matched the single-value forecast temperature. The last two nights in each verification set, for which participants had seen the forecast, resulted in correct rejections. Both the forecast temperature and the observation were above 32°F. Each participant saw two different forecasts: one from weather dataset A and the other from weather dataset B. Half of participants saw forecast A first and the other half saw forecast B first.

[2] Two weeks was selected as the maximum length for the graphic that was clearly visible in most web browsers, but long enough to depict trends.

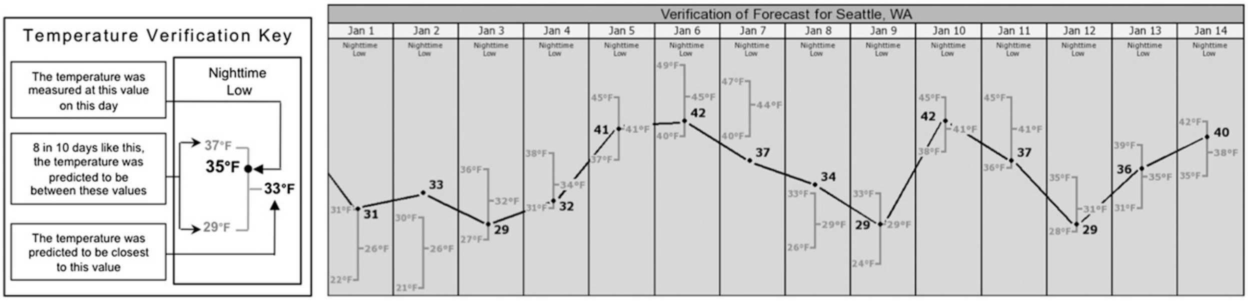

FIG. 4. Predictive interval forecast verification graphic for experiment 1.

In the predictive interval condition in which verification was provided, participants saw one example of each of the calibration levels from different original weather datasets. Half the participants saw a 79% calibration first and half saw a 64% calibration first. Thus there were three verification conditions with approximately equal numbers of participants: well-calibrated (79%) first, ill-calibrated (64%) first, and no-verification. The forecasts and observations in the deterministic condition that corresponded to these calibration sets were similarly counterbalanced.

5) GRAPHICS

The single-value forecast was represented by a black point and number, located vertically in a box with higher position indicating higher temperature, 10 pixels for every one degree Fahrenheit (Fig. 1). The key read, ‘‘The temperature will usually be closest to this value.’’ The predictive interval included a bracket as well, with numbers at the ends indicating the upper and lower bound temperatures of the 80% predictive interval (Fig. 2). The predictive interval key read ‘‘8 in 10 days like this, the temperature will be between these values.’’ A frequency definition was used because it corresponded directly to the calibration metric and has been shown to be superior to probability in some reasoning tasks (e.g., Gigerenzer and Hoffrage 1995).

The verification graphics were built on the forecast graphics for ease of interpretation (Figs. 3 and 4). They showed 14 nighttime low temperature forecasts in chronological order from left to right with forecasted values in dark gray and observed values as black dots connected with a black line. The verification keys included an additional black arrow pointing to observed temperature that said, ‘‘The temperature was measured at this value on this day.’’

b. Results

First we examined participants’ temperature estimates to determine whether they understood the forecast. Then, to determine whether the predictive interval influenced participants’ response to the forecast, we examined freeze warning decisions, range of temperature expectations, and certainty ratings comparing the predictive interval to the deterministic format conditions. Next we examined answers to questions targeting participants’ understanding of the verification graphics. Finally we examined whether participants would choose to use the forecast again to determine whether predictive intervals or verification influenced trust in the forecast.

1) UNDERSTANDING THE FORECAST

Some participants in the predictive interval condition appeared to misunderstand the temperature forecast. Asked what they thought the daytime high temperature would be, they gave the value at the upper bound of the interval. For the nighttime low temperature they gave the value at the lower bound of the interval. There were two possible explanations for these ‘‘errors.’’ One was that participants mistook the bracket for an expression of diurnal fluctuation with the top indicating the high temperature and the bottom the low temperature for each 12-h period. Alternatively, participants may not have believed that the forecast would verify exactly and choose these values coincidentally. To distinguish between these two explanations, we compared the predictive interval condition to the deterministic condition where no boundary values were included. For each participant we calculated the total errors over the four temperature questions in the first forecast and the four temperature questions in the second forecast. Indeed, there were significantly more errors in the predictive interval than in the deterministic condition, ruling out the ‘‘coincidence’’ explanation (Table 1). A repeated-measures one-way

TABLE 1. Mean number of deterministic construal errors in experiment 1 with standard deviations in parentheses.

Forecast format	Forecast A	Forecast B	Total	*N*
Predictive interval	0.82 (1.18)	0.48 (0.98)	0.65 (0.72)	155
Deterministic	0.20 (0.48)	0.29 (0.58)	0.25 (0.72)	147
Total	0.51 (0.63)	0.39 (0.59)	0.45 (0.71)	302

TABLE 2. Percent of participants issuing a freeze warning for each single-value temperature forecast in experiment 1 (excluding those who made DCE on that forecast).

Temperature forecast	33°F	35°F	37°F	38°F
Predictive interval $N = 155$	74%	23%	8%	3%
Deterministic $N = 147$	67%	23%	4%	4%
Total $N = 302$	71%	25%	6%	3%

TABLE 3. Mean certainty rating selected by participants in experiment 1 for temperatures above the upper bound, between the two bounds, and below the lower bound (standard deviations in parentheses).

	Above	Between	Below	N
Predictive interval	28% (19%)	75% (16%)	24% (16%)	129
Deterministic	35% (16%)	75% (17%)	31% (16%)	104

analysis of variance (ANOVA) was conducted on mean errors with forecast format (predictive interval, deterministic) as the between groups independent variable and forecast order (forecast 1, forecast 2) as the within groups independent variable. Participants with predictive interval forecasts made significantly more errors than did those with deterministic forecasts, $F(1, 300) = 23.41$, $p < 0.001$. However, there were significantly fewer errors on the second as compared to the first forecast, $F(1, 300) = 5.41$, $p = 0.02$. In addition, there was a significant interaction between forecast format and forecast order, $F(1, 300) = 15.60$, $p < 0.001$. Errors decreased in the predictive interval condition from the first to the second forecast, although they remained low and approximately the same in the deterministic condition. Because this suggests that participants using the predictive interval misconstrued it as a deterministic forecast with additional information about diurnal fluctuation, we refer to it as a ‘‘deterministic construal error’’ (DCE). Although the DCE constitutes a serious misinterpretation, it appears that prior exposure alone reduces it significantly, without explicit explanation.

2) Influence of the predictive interval

The goal of the task was to decide whether to post a freeze warning. Few participants in any condition decided to post a warning for the single-value forecasts of 35°, 37°, and 38°F (Table 2) and there were no significant differences by forecast format (35°F, $p = 0.36$; 37°F, $p = 0.19$; and 38°F, $p = 0.47$). We focus here on the forecast for 33°F, for which 71% of participants overall decided to post a warning. We conducted a logistic regression analysis on the binary decision with forecast format (predicative interval, deterministic) and forecast order (forecast 1, forecast 2) as the independent variables. A significantly greater proportion of those in the predictive interval than in the deterministic condition decided to post a warning, $\mathrm{Exp}(B) = 13.74$, $p = 0.003$. In addition, participants were more likely to post a warning when the 33°F forecast was presented second (81%) than first, (52%), $\mathrm{Exp}(B) = 8.31$, $p < 0.001$. Moreover, there was a significant forecast order by forecast format interaction, $\mathrm{Exp}(B) = 4.35$, $p < 0.007$. Few of those in the deterministic condition posted a warning for 33°F in the first forecast (38%) while many more did so for 33°F in the second forecast (84%). For those in the predictive interval condition, however, the majority posted warnings for 33°F in both the first (66%) and second (79%) forecast.

The impact of the predictive interval was also seen in participants' temperature expectations. Participants with predictive intervals expected lower temperatures overall than did those using the deterministic forecast. Participants indicated the temperatures they would not be surprised to observe above and below each of the four nighttime low temperature forecasts. We subtracted participants' answers from the single-value forecast and calculated a mean difference score for the high-end and low-end estimates across all four questions. A multivariate ANOVA conducted on participants' mean high- and low-end difference scores with forecast format (predictive interval, deterministic) as the independent variable revealed that the mean low-end difference score was significantly smaller in the deterministic condition ($M = -2.86$, SD = 1.43) than in the predictive interval condition ($M = -3.92$, SD = 1.13), $F(1, 300) = 51.44$, $p < 0.001$. The mean high-end difference scores in the predictive interval ($M = 3.72$, SD = 1.85) and deterministic conditions ($M = 3.69$, SD = 2.55) were not significantly different. This suggests that those with the predictive interval anticipated lower temperatures than did those using deterministic forecasts.

The influence of the predictive interval was also obvious in participants' certainty ratings made after the second forecast (Table 3).[3] Participants in the predictive interval condition expected a significantly smaller chance of temperatures a few degrees from the forecasted value. A multivariate ANOVA was conducted on mean percent chance selected for above the upper bound temperature, below the lower bound temperature, and between the two, with forecast format (predictive interval, deterministic) as the independent variable. Participants with predictive interval forecasts

[3] Sixty-six participants, evenly distributed across conditions, were omitted from this analysis because of an error in the question. Therefore the N for these analyses is 231.

TABLE 4. Mean forecast performance rating (standard deviations in parentheses) for first and second verification period by calibration in experiment 1 (7-point scale with higher number indicating better performance).

	First verification period			Second verification period		
Forecast format	79%	64%	Total	79%	64%	Total
Predictive interval	4.39 (1.19) $N = 54$	4.58 (0.93) $N = 57$	4.49 (1.06) $N = 111$	4.86 (1.13) $N = 57$	3.65 (1.00) $N = 54$	4.27 (1.22) $N = 111$
Deterministic	4.45 (0.86) $N = 55$	4.26 (0.91) $N = 54$	4.36 (0.89) $N = 109$	4.61 (1.19) $N = 54$	4.07 (1.14) $N = 55$	4.34 (1.19) $N = 109$
Total	4.42 (1.03) $N = 109$	4.42 (0.93) $N = 111$	4.42 (0.98) $N = 220$	4.74 (1.16) $N = 111$	3.87 (1.08) $N = 109$	4.30 (1.20) $N = 220$

indicated a significantly smaller chance of observing temperatures above the upper bound than did those with deterministic forecasts, $F(1, 229) = 11.24$, $p = 0.001$. They indicated a significantly smaller chance of observing temperatures below the lower bound than did those with deterministic forecasts, $F(1, 229) = 10.27$, $p = 0.002$. However, the mean percent chance of observing temperatures between the two boundary values was similar for the two forecast formats and not significantly different, $F(1, 229) = 0.02$, $p = 0.88$.

Taken together, this set of analyses demonstrates that the 80% predictive interval narrowed the range of temperatures participants considered likely and alerted them to colder outcomes. This combination may seem initially contradictory. However, it is important to remember that narrowed range expectations were revealed in smaller *probability estimates* for observing temperatures a few degrees from the single-value forecast. Colder outcome expectations were revealed in lower *temperature values,* indicating what would not surprise participants. We do not know the probability that corresponds with participants' notion of ''not surprising,'' so it is not possible to match up participants' answers to these two different questions. In fact, it may be that predictive interval changes what would not surprise people, precisely because it informs them of a range of possible values. The bottom line, however, is that this combination allowed them to post the warning more often for the 33°F temperature forecast.

3) UNDERSTANDING VERIFICATION GRAPHICS: EVALUATING FORECAST PERFORMANCE

The next set of analyses, conducted on participants' judgments of forecast performance over various time periods, tested whether they understood the verification graphics. Participants in the 80% predictive interval condition saw two verification periods, one that was well calibrated (79%) and one that was not (64%). If participants understood the notion of calibration based on the graphic alone, ratings for the 79% period should be significantly higher than ratings for the 64% period and that difference should exceed any difference in rating among those in the deterministic condition who were not exposed to calibration differences. For the first verification period, all of the mean ratings[4] were just above the middle of the 7-point scale (Table 4) and there were no significant differences. At that point participants had nothing with which to compare, having seen only one verification graphic. However, for the second verification period a clear pattern emerged in which participants with the predictive interval rated the 79% calibration higher and the 64% calibration lower than did those with the deterministic forecast. A univariate ANOVA conducted on performance ratings in the second verification period, with forecast format (predictive interval, deterministic) and forecast calibration (79%, 64%) as the between-groups independent variables revealed that participants rated the 79% calibration significantly higher than the 64% calibration, $F(1, 216) = 33.95$, $p < 0.001$. In addition, there was a significant interaction between forecast calibration and forecast format $F(1, 216) = 5.02$, $p = 0.026$. Those in the predictive interval condition made a greater distinction between the two calibrations than did those with the deterministic forecast.

We also asked participants to identify the single week within each 2-week period in which the forecast performed better. Because the second week in each period had both lower error and better calibration, participants who understood the verification graphic in both the deterministic and the predictive interval conditions should have chosen the second week. Indeed, this was the choice of the majority of participants in both forecast format conditions (Table 5). Chi-square analyses conducted on the number of participants selecting the correct week in the first verification period confirmed that the proportion was significantly greater than would be

[4] The original rating scale ranged from 1 (outstanding) to 7 (terrible); however, it was inverted in these analyses so that a higher rating indicates better performance.

TABLE 5. Percent of participants in experiment 1 identifying second week forecast, with less error and better calibration, as better performing.

Verification period	First	Second	Total
Predictive interval	87%	87%	87% ($N = 111$)
Deterministic	92%	86%	89% ($N = 109$)
Total	89%	87%	88% ($N = 220$)

TABLE 6. Mean and standard deviation of forecast performance ratings (1–3 with higher number indicating better performance) for three individual days in experiment 1.

	Within day	Outside day	Matching day	Total
Predictive interval	2.15 (0.58)	1.29 (0.43)	2.91 (0.26)	2.11 (0.27) $N = 111$
Deterministic	1.42 (0.47)	1.44 (0.44)	2.9 (0.37)	1.93 (0.27) $N = 109$
Total	1.79 (0.64)	1.36 (0.44)	2.9 (0.27)	$N = 220$

expected by chance in both the deterministic ($\chi^2 = 75.97$, $p < 0.001$) and the predictive interval conditions ($\chi^2 = 59.11$, $p < 0.001$). Similarly, for the second verification period, significantly more participants than expected by chance selected the correct week in both the deterministic ($\chi^2 = 57.26$, $p < 0.001$) and the predictive interval conditions ($\chi^2 = 62.06$, $p < 0.001$).

Finally, participants rated forecast performance on three individual days. We focus here on the two days with identical error for which one observation was within and the other outside of the predictive interval bracket. If those in the predictive interval condition understood that the intent of the bracket graphic was to capture the majority of observed temperatures, a concept that was not explicitly explained to participants, they should have rated the ''within'' day higher than the ''outside'' day. Indeed, the majority of participants with the predictive interval did so. However, those in the deterministic condition rated the two days approximately equally (Table 6). For this analysis, the patterns of ratings in the first and second verification periods were similar so we combined them by calculating means for each day. A repeated measure ANOVA on ratings, with days (''within'', ''outside'') as the within-groups independent variable and forecast format (predictive interval, deterministic) as the between-groups independent variable revealed that participants in the predictive interval condition rated the days significantly higher overall than did those in the deterministic condition $F(1, 218) = 14.85$, $p < 0.001$. In addition, the ''within'' day was rated significantly higher than the ''outside'' day, $F(1, 218) = 93.88$, $p < 0.001$. Importantly, there was a significant interaction between forecast format and day indicating that there was a greater discrepancy in the rating for the two days among participants in the predictive interval condition than in the deterministic condition, $F(1, 218) = 113.78$, $p < 0.001$. In other words, although the error in the single-value forecast was identical, those in the predictive interval regarded performance of the two forecasts as markedly different.

It is obvious from these analyses that participants in both conditions understood the verification graphics based on the written explanation in the key. This information, pointing out the forecasted and observed temperatures, was sufficient to allow those in the deterministic condition to select the week with the least error. It was sufficient to allow those in the predictive interval condition to identify the two-week period that was well calibrated, despite the fact that the single-value forecast error was almost identical to that in the ill-calibrated period. Furthermore it is clear that the majority of participants understood the intent of the predictive interval graphic because they rated the day in which the observation was within the bracket more highly than the one in which it was outside of the bracket.

4) TRUST IN THE FORECAST

One of our main questions concerned trust in the forecast. As an indication of trust, we asked participants if they would use the same kind of forecast again. A greater proportion of those with the predictive interval forecast said ''yes'' both times they were asked (Table 7). Two separate binary logistic regressions were performed, one for participants' responses to this question asked after each forecast. After the first forecast, participants in the predictive interval condition were more than twice as likely to decide to use the forecast again as were those in the deterministic condition, $\mathrm{Exp}(B) = 2.12$, $p = 0.002$. After the second forecast, participants in the predictive interval condition were again more than twice as likely to decide to use the forecast again as those in deterministic condition, $\mathrm{Exp}(B) = 2.31$, $p < 0.001$. There was no effect of verification, $\mathrm{Exp}(B) = 0.74$, $p = 0.25$, tested only in the second forecast. Interestingly, a McNemar's chi-square statistic suggests that significantly fewer participants overall said they would use the forecast again the second time they are asked as compared to the first time they are asked, $\lambda^2 = 27.91$, $p < 0.001$.

c. *Discussion of experiment 1*

Thus, experiment 1 established that the majority of participants understood both the predictive interval and verification graphics based on a simple key. Moreover,

TABLE 7. Percent of participants in experiment 1 indicating that they would use the forecast again.

		Forecast 2		
Forecast format	Forecast 1 (no verification)	No verification	Verification (64% and 79%)	Total
Predictive interval	66% $N = 155$	57% $N = 44$	51% $N = 111$	53% $N = 155$
Deterministic	48% $N = 147$	39% $N = 38$	30% $N = 109$	33% $N = 147$
Total	58% $N = 302$	49% $N = 82$	41% $N = 220$	43% $N = 302$x

predictive interval graphics significantly influenced expectations, 1) narrowing the range of temperatures that seemed likely, thereby 2) allowing participants to consider lower temperatures and 3) encouraging them to take precautionary action more often.

To understand these advantages, it is useful to consider responses in the deterministic condition. Notice that those in the deterministic condition expected almost a degree less error below as opposed to above the single-value forecast. This may be because they regarded colder temperatures as locally unusual. Indeed, there is evidence that people expect rare single-value forecasts to verify closer to normal values, in this case, warmer (Joslyn and Savelli 2010). Perhaps this effect extends to the merely unusual temperatures of concern here. If so, it suggests that predictive intervals may assist people in overcoming prior expectations, in this case by alerting them to the real possibility of colder temperatures in the current situation. This in turn encourages them to precautionary action.

It is also interesting to note that in the deterministic condition the majority of participants posted warnings for the 33°F forecast only when it was presented *second*, suggesting that they needed to observe several forecasts to judge when precautionary action was warranted. However, the majority of those in the predictive interval condition posted a warning for 33°F regardless of whether it was presented first or second, suggesting that the predictive interval supplied the relevant information.

Thus, experiment 1 provides evidence for several clear advantages for predictive interval forecasts that could extend to many similar decision-making tasks. Importantly, although the predictive interval graphic was substantially more complex, a significantly greater proportion of participants claimed they would use it again as compared to those using the deterministic forecast, demonstrating that participants were aware of the value of the predictive interval forecast and suggesting that they trusted it.

However, some of those using the predictive interval initially misinterpreted the graphic. Their responses suggested that they thought it was a deterministic forecast with additional information about diurnal fluctuation. This is psychologically similar to errors in interpretation for other probabilistic forecasts. For instance, a common misinterpretation of probability of precipitation is that it is a deterministic forecast with additional information about the proportion of time or area over which precipitation will be observed (Murphy et al. 1980; Gigerenzer et al. 2005; Joslyn et al. 2009). There is anecdotal evidence for a similar misinterpretation of the cone of uncertainty, showing potential hurricane tracks. It is often misinterpreted as depicting the wind field, a deterministic forecast. Why do people make this mistake? Misinterpreting probabilistic forecasts as deterministic may function to reduce cognitive load. A probabilistic forecast requires that one continues to consider multiple possible outcomes throughout the decision process whereas a deterministic forecast does not. Thus, the deterministic interpretation may be psychologically ''easier,'' although this is not necessarily a conscious choice. Some participants may have simply assumed that the forecast expressed diurnal fluctuation, based on the graphics alone without bothering to read the key. This may have been exacerbated by the lengthy definition provided for the single-value forecast. These issues were addressed in experiment 2.

In addition, although those in the predictive interval condition expected a smaller chance of observing temperatures beyond the interval than did those in the deterministic condition, they expected almost twice as great a chance of observations above and below the interval (about 20%) as was intended by the forecast (10%). This is what one would expect if participants regarded the forecast as overconfident. It suggests that participants believed the interval would fail to capture as many observations as intended. On the other hand this effect may have been due to the definition provided in the key that focused on the values within the interval (8 in 10). To determine the chance of observations above or below the interval, a multiple step calculation was required involving converting frequency into percent [$(8/10) \times 100 = 80\%$], subtracting that amount from 100% ($100\% - 80\%$), and then dividing by 2 ($20\%/2 = 10\%$).

Omitting the last step in this process could lead to the error that was observed here. We attempted to resolve this issue in experiment 2 as well.

Finally, although it was clear that participants understood the verification graphics, verification did not impact trust in the forecast. However, effects on trust may have been obscured by effects due to forecast order. Recall that there was a reduction in the number of participants saying they would use the forecast again after the second forecast. Because the impact of verification on trust could only be tested in the second forecast after participants had been introduced to the verification graphics the order effect may have overpowered an effect of verification. Experiment 2 attempted to disentangle the two.

3. Experiment 2

Experiment 2 was conducted to replicate the main findings of experiment 1, to test methods for reducing the two errors in interpreting the predictive interval forecast described above, and to provide a stronger test of the effect of verification on user trust.

In an attempt to reduce deterministic construal errors, the experimenter read the keys aloud to ensure that all participants were exposed to this information at the beginning of the experiment. In addition, the definition for the single-value forecast was simplified to ''best forecast'' to emphasize this value. If the DCEs in experiment 1 were due to ignoring or misunderstanding the definition provided in the key, then there should be no more DCEs in the predictive interval than in the deterministic conditions in experiment 2.

In addition, the predictive interval definition was changed to specify the *probability* (10%) of observing temperatures *beyond* each boundary. If the probability selected by participants in experiment 1 was due to a calculation error, rather than to regarding the forecast as overconfident, then it should be reduced with the new definition.

The procedure in experiment 2 was simplified as well in an attempt to disentangle any potential verification effects from order effects. Participants responded to a single forecast presented after an instruction phase in which all of the graphics were introduced. This allowed us to test whether exposure to the verification graphic, manipulated in the instruction phase, affected participants trust in their initial encounter with the forecast.

a. *Method*

1) Design

The experiment was a two by two full factorial design. Participants were randomly assigned to either a predictive interval or a deterministic forecast condition. Within each forecast format condition participants were randomly assigned to either verification or no verification. The dependant variables were the expected temperatures, freeze warning decision, uncertainty rating, forecast performance evaluation, and trust decision.

2) Participants

Participants were 312 University of Washington introductory psychology students of whom 54% were women. They ranged in age from 18 to 39 years old. They received course credit for their participation.

3) Procedure

The procedure for experiment 2 was similar to that of experiment 1 with two exceptions. First, after explaining the computer interface and the freeze warning task, the experimenter showed an example of the forecast graphic and read the text in the key aloud (see the appendix). For those in the verification conditions, the experimenter then showed a 14-day verification graphic, with 79% of the observations within the interval, and read the key aloud. After the instruction phase, participants answered the same questions as were asked in experiment 1 about a single 2-day forecast. Uncertainty ratings were requested only of participants in the nonverification conditions. Those in the verification conditions were not exposed to the uncertainty concepts introduced in the questions to avoid influencing their understanding of the verification graphic. Instead, a verification graphic was presented and participants answered the three forecast performance evaluation questions from experiment 1. The rating scale for the question targeting 2-week period and individual days was changed to 7-point scale, ranging from 1 (very bad) to 7 (very good), to make them more similar to one another.

4) Weather data

The forecasts and observations were based on the same archived forecast data used in experiment 1. The first 14-day period, used in the instruction phase, was adjusted so that 79% of the observations were within the interval and the standard error was 3.9°F. The second 14-day period was adjusted to produce two calibration levels, 79% (SE = 3.5) and 64% (SE = 3.7). For both verification sets the mean error above and below the deterministic forecast was about 3.5°F and mean predictive interval width was 7.71°F. The predictive interval widths for three of the four forecasts (2 days and 2 nights) were reduced to 6°F to provide a more precise forecast. Again, there were three verification conditions with approximately equal numbers of participants: well-calibrated (79%), ill-calibrated (64%), and no-verification.

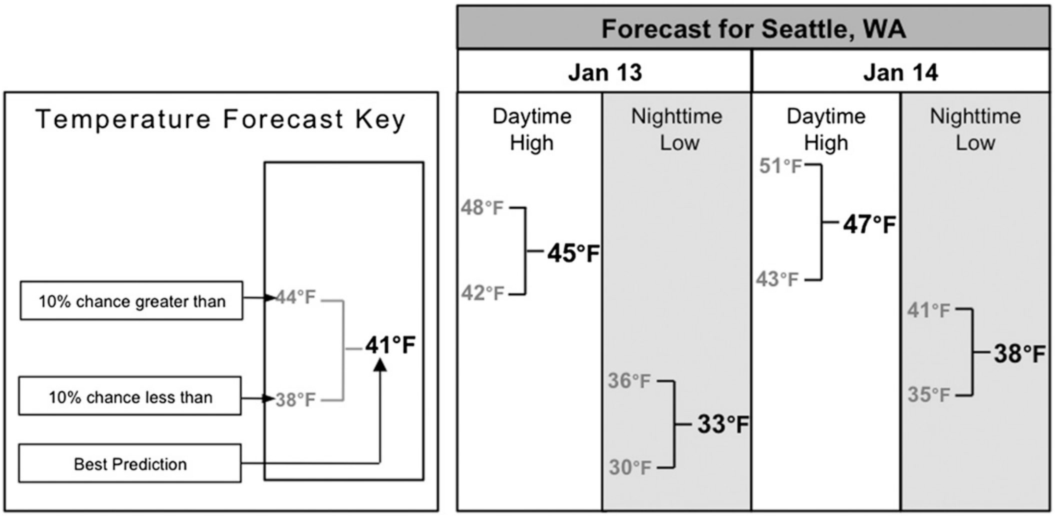

FIG. 5. Predictive interval forecast graphic for experiment 2, forecast A.

5) GRAPHICS

The graphics from experiment 1 were changed for experiment 2 to reflect the changes in forecast values and text in the key described above (Fig. 5).

b. Results

1) UNDERSTANDING THE FORECAST

Using the same data analysis plan as for experiment 1, we began by examining participants' temperature estimates. Despite reading the key aloud and simplifying the definition, DCEs were approximately as common in the predictive interval condition in experiment 2 ($M = 0.63$, SD = 1.04) as they had been in experiment 1 ($M = 0.65$, SD = 0.72). Moreover, the mean number of DCEs across all four questions was significantly greater in the predictive interval than in the deterministic forecast condition ($M = 0.20$, SD = 0.47) according to an independent samples t test, $t(1, 310) = -4.75$, $p < 0.001$.

2) IMPACT OF THE PREDICTIVE INTERVAL

As with experiment 1, significantly more participants with predictive intervals (60%) than with deterministic forecasts (41%) decided to post a freeze warning for the 33°F forecast, Exp(B) = 2.18, $p = 0.001$. Very few participants (2%) decided to post a warning for the warmer (38°F) nighttime low forecast and there were no significant differences between groups, Exp(B) = 5.38, $p = 0.13$.

Participants' temperature expectations were also influenced by predictive interval forecasts. As with experiment 1, those using predictive intervals indicated that they would not be surprised by temperatures that were generally colder than temperatures indicated by participants using deterministic forecasts. Again, difference scores were calculated by subtracting participants' high- and low-end temperature estimates from the single-value forecast. A multivariate ANOVA was conducted on mean high-end and low-end difference scores with forecast format (predictive interval, deterministic) as the independent variable. Participants' mean low-end difference score was significantly larger in the predictive interval condition ($M = -3.41$; SD = 1.71) than in the deterministic condition ($M = -2.54$; SD = 2.01), $F(1, 310) = 16.92$, $p < 0.001$. Participants' mean high-end difference score was significantly smaller in the predictive interval condition ($M = 3.38$; SD = 1.72) than in the deterministic condition ($M = 4.08$; SD = 2.53), $F(1, 310) = 8.04$, $p = 0.005$.

The influence of the predictive interval was also obvious in participants' certainty ratings (Table 8). As with experiment 1, participants in the predictive interval condition expected a significantly smaller chance of temperatures a few degrees from the forecasted value.

TABLE 8. Mean certainty rating selected by participants for observing temperatures above the upper bound, between the two bounds, and below the lower bound in experiment 1 (no-verification condition only) and experiment 2 (standard deviations in parentheses).

Forecast format		Above	Between	Below
Predictive interval	Experiment 1 $N = 44$	22% (16%)	81% (13%)	20% (15%)
	Experiment 2 $N = 50$	13% (11%)	78% (16%)	14% (13%)
Deterministic	Experiment 1 $N = 38$	33% (15%)	79% (19%)	24% (14%)
	Experiment 2 $N = 50$	29% (18%)	74% (21%)	33% (20%)

TABLE 9. Mean forecast performance ratings (1–7) for the three individual days for experiment 2 with standard deviation in parentheses.

Forecast format	Within day	Outside day	Matching day	Total
Deterministic $N = 109$	3.38 (1.2)	3.51 (1.11)	6.22 (0.73)	4.37 (0.08)
Predictive interval $N = 103$	4.83 (1.18)	2.91 (1.25)	6.21 (0.74)	4.65 (0.08)
Total $N = 212$	4.17 (0.77)	3.15 (1.25)	6.22 (0.73)	4.51 (0.06)

A multivariate ANOVA was conducted on mean percent chance selected above the upper bound temperature, below the lower bound temperature, and between the two, with forecast format (predictive interval, deterministic) as the independent variable. Participants with predictive interval forecasts indicated a significantly smaller chance of observing temperatures above the upper bound than did those with deterministic forecasts $F(1, 98) = 26.68$, $p < 0.001$. They indicated a significantly smaller chance of observing temperatures below the lower bound than did those with deterministic forecasts, $F(1, 98) = 32.62$, $p < 0.001$. Although the mean percent chance of observing temperatures between the two boundary values was larger in the predictive interval than deterministic condition, $F(1, 98) = 1.56$, $p = .24$, the difference did not reach significance.

Notice that the mean percent chance in the deterministic condition is remarkably similar to that in experiment 1; however, in the predictive interval condition, the percent chance beyond the boundary values is about half a large as that in experiment 1 and much closer to the 10% intended by the forecast, suggesting that the change in key definition helped. Indeed, this difference was significant in a multivariate ANOVA[5] conducted on the three intervals (i.e., above the upper bound, below the lower bound, and between the two boundaries) with experiment (1, 2) and forecast format (predictive interval, deterministic) as independent variables. The mean percent chance above the upper bound was significantly smaller in experiment 2, $F(1, 178) = 8.24$, $p = 0.005$. For mean percent chance below the lower bound, there was a significant interaction $F(1, 178) = 10.10$, $p = 0.002$, indicating that the mean percent chance selected in the deterministic condition was larger in experiment 2 while it was smaller in the predictive interval condition.

3) UNDERSTANDING VERIFICATION GRAPHICS: EVALUATING FORECAST PERFORMANCE

Next we examined participants' evaluation of forecast performance over the three time periods. For the two-week period, those in the predictive interval condition rated the 79% calibration higher ($M = 4.39$, SD = 1.33) than the 64% calibration ($M = 4.02$, SD = 1.18). Those in the deterministic condition rated the 79% ($M = 4.29$, SD = 1.44) only slightly higher than the 64% ($M = 4.17$, SD = 1.12). However, neither the main effect for calibration $F(1, 208) = 1.85$, $p < 0.18$ nor the interaction $F(1, 208) = 0.49$, $p < 0.48$ reached significance in this analysis. For the one-week period, chi-square analyses revealed that a significant majority of those in both the deterministic (98%), $\chi^2 = 101.15$, $p < 0.001$, and predictive interval (92%), $\chi^2 = 70.14$, $p < 0.001$ conditions correctly choose the week with lower error and better calibration. For the individual day ratings (Table 9), a repeated measures ANOVA was conducted on rating for the two days with identical error, with day (observation within, outside the predictive interval) as the within-groups independent variable and forecast format (deterministic, predictive interval) as the between-groups independent variable. Participants with the predictive interval rated the days significantly higher overall than did those with the deterministic forecast $F(1, 210) = 9.34$, $p = 0.003$. In addition, participants rated the ''within'' day significantly higher than the ''outside'' day, $F(1, 210) = 152.20$, $p < 0.001$. Importantly, there was a significant interaction between forecast format and day, indicating that the discrepancy in rating between the two days was much greater for participants in the predictive interval condition than for those in the deterministic condition, $F(1, 210) = 114.08$, $p < 0.001$.

4) TRUST IN THE FORECAST

Finally, we asked how the graphics influenced participants' trust in the forecast as reflected in their choice to use the same kind of forecast again. Here we found that trust was enhanced both by the predictive interval forecast and by verification (Table 10). A binary logistic regression was performed on participants response with forecast format (deterministic, predictive interval) and verification (verification, no verification) as the between-participants independent variables. Participants in the predictive interval condition were almost twice as likely to decide to use the forecast again as were those in the deterministic condition, Exp(B) = 1.75, $p = 0.03$. In addition, participants with verification were twice as likely to decide to use the forecast again as those without verification Exp(B) = 2.18, $p = 0.003$.

[5] We include only the no-verification condition from experiment 1 so that it is comparable to experiment 2 in which these questions were only asked in the no-verification conditions.

TABLE 10. Percent of participants indicating that they would use the forecast again in experiment 2.

Forecast format	No verification	Verification	Total
Deterministic	46% $N = 50$	73% $N = 109$	65% $N = 159$
Predictive interval	72% $N = 50$	78% $N = 103$	76% $N = 153$
Total	59% $N = 100$	75% $N = 212$	70% $N = 312$

c. *Discussion of experiment 2*

All of the major effects found in experiment 1 were replicated in experiment 2, providing strong evidence that predictive interval forecasts influence participants understanding of future weather events as well as the resulting decisions. As with experiment 1, participants in both conditions appear to have understood the verification graphics. Both those using the deterministic and predictive interval forecasts selected the week with the least error. Predictive interval participants rated the day for which the observation fell within the interval significantly higher than the day for which it fell outside of the interval, despite the fact that both observations were 4°F from the single-value forecast. In addition, most of those in the predictive interval condition identified the better-calibrated 2-week period, although the effect failed to reach significance perhaps because of the simplified procedure in which only one verification graphic was evaluated. Recall that in experiment 1 the effect was prominent only on the second evaluation after participants had interacted with a previous graphic.

Only one of the two attempts to reduce interpretation errors was completely successful. Participants estimated smaller percentages beyond the predictive interval with the new probability-beyond definition, suggesting that much of the overestimation in experiment 1 was due to a calculation error. Unfortunately, however, the number of DCEs was approximately equivalent to that observed in experiment 1, despite reading the key out loud and referring to the single-value forecast as ''best.'' This suggests that for some participants the DCE is a particularly persistent misconception. The implications of these two results will be discussed in the conclusions below.

In addition, in experiment 2 there was a significant advantage for verification. Both predictive interval and verification graphics appeared to have increased trust in the forecast, as those who used them were more likely to choose to use the same kind of forecast again than were those with no verification graphics.

4. Conclusions

In the two experiments reported here, we demonstrated that nonexpert end users understood simple graphics depicting the 80% predictive interval as well as verification graphics for both deterministic and probabilistic forecasts. Because participants were given only the information in a simple key and no explicit training, this suggests that such information could be successfully conveyed to general public end users in a web or print format.

Moreover, the 80% predictive interval significantly influenced participants' understanding of the future weather. In this freeze warning task, it alerted users to the possibility of colder temperatures, at the same time narrowing the range of temperatures that were regarded as likely. Importantly, the predictive interval influenced decisions, encouraging participants to take precautionary action. As such, the predictive interval forecast reduced risk seeking, an error that is common in situations in which precautionary action is regarded as costly (Joslyn and LeClerc 2012).

Furthermore, the predictive interval increased trust in the forecast. Those using the predictive interval were more likely to say that they would choose it again, in both experiments, compared to those using the deterministic forecast. This is particularly impressive because the predictive interval forecast was considerably more complex, with more information to process, much of which was abstract in nature and unfamiliar to the undergraduate participants in this study. These results suggest that, despite the additional cognitive load, participants clearly saw the value of the predictive interval forecast.

What about the impact of verification? In experiment 1 trust in the forecast was not influenced by verification graphics, perhaps because of the order effects detected in that experiment. When the procedure was simplified in experiment 2, those with verification were more likely to say that they would use the forecast again, constituting evidence that verification also increased trust in the forecast. This was true both of the predictive interval and deterministic conditions, suggesting that simple verification graphics may provide general public end users with information required to evaluate a wide range of forecasts. However, the fact that this effect was obscured by order in experiment 1 suggests that it is not robust. Obviously additional research is required before clear recommendations can be made.

We also noted two important misunderstandings of the graphics as well as methods for reducing them. In experiment 1 participants thought the probability of observations beyond the 80% predictive interval was about twice what was intended. This appears to have been largely due to the definition, which focused on the proportion of observations expected *within* the interval, thus requiring additional calculations to determine the percentage beyond it. In experiment 2, the interval was defined in terms of the probabilities of observing temperatures above and below the boundaries of the interval, providing participants with information that was compatible with the questions targeting these values. This led to probability estimates that were much more in line with what was intended and illustrates the importance of expressing the forecast in a manner that is compatible with use. Interestingly, the ''beyond'' definition used in experiment 2 did not appear to have a negative impact on participants understanding of the probability of observing temperatures *within* the interval. In fact, the estimate was somewhat closer to the 80% intended by the predictive interval forecast in experiment 2 than it was in experiment 1 in which the *within* definition was used. This suggests that the ''probability-beyond'' definition used in experiment 2 may indeed be more easily understood by users than the ''frequency-within'' definition that was used in experiment 1. In addition, it appears that there was little difference between the two definitions in all other respects, as the results were very similar for the two experiments in terms of both understanding the predictive interval and its impact on decision making.

The second major misunderstanding was the deterministic construal error. Not only does this error prevent users from taking advantage of the uncertainty information, but it also affects their understanding of the single-value forecast. The good news is that DCEs were considerably reduced by prior experience and familiarity with the forecast. In experiment 1 we saw a significant decline in DCEs on the second forecast. We did *not* see an equally low level of errors with mere exposure to the forecast in the instruction phase of experiment 2. Perhaps one must deliberately interact with the forecast, using it to answer specific questions, to reach this level of understanding.

Since these misunderstandings have clear solutions we are confident that predictive intervals could be beneficial to general public end users in a variety of domains. The predictive interval is a particularly adaptable form of uncertainty information because the forecaster need not know the users' specific threshold of concern. As such the same forecast can be applied to a variety of decision tasks. Furthermore, because predictive intervals explicitly acknowledge forecast uncertainty, they can be regarded as reliable when compared with a range of observations. Thus, forecasts that include an estimate of uncertainty, like the predictive interval, may well counteract the negative impact of high-profile misses such as ''No-maggeden'' described in the introduction.

Acknowledgments. This material is based upon work supported by the National Science Foundation Grant 1023354.

APPENDIX

Instructions to Participants

a. Experiment 1

''These are weather forecasts from a new weather website. Please answer the questions below the display to the best of your ability, using the information on the website. Assume that it is your job to issue freeze warnings (temperatures are expected to fall below 32°F) in an agricultural community. Freezing temperatures can result in partial loss in yield and/or quality of current and future harvests. Farmers use the warning to decide when to protect crops. Protective measures involve covering plants with lightweight plastic to reduce heat loss. These measures are costly, as additional labor must be hired to place the covers over the plants and then remove them when temperatures rise above freezing. For this reason, it is important NOT to issue freeze warnings unless freezing temperatures are expected, because farmers would implement expensive precautions for which there was no need.''

b. Experiment 2

Forecast graphic is shown and experimenter says, ''These are weather forecasts from a new weather website. Please study the graphic and be sure you understand it. Notice the forecast is for two days, with nighttime low and daytime high information for both days. There is also a key below. The key says [experiment reads text in key].''

Verification graphic is shown and experimenter says, ''Below is a two-week graph for the nighttime low temperature from January 1st to 14th that shows how well the forecast did. Please study the graphic and be sure you understand it. Again, there is a key below. It says [experimenter reads text in key].''

[Experimenter then reads task description identical to experiment 1 instructions here.]

REFERENCES

Alpert, M., and H. Raiffa, 1982: A progress report on the training of probability assessors. *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, Eds., Cambridge University Press, 294–305.

Erev, I., T. S. Wallsten, and D. V. Budescu, 1994: Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychol. Rev.,* **101,** 519–527.

Gigerenzer, G., and U. Hoffrage, 1995: How to improve Bayesian reasoning without instruction: Frequency formats. *Psychol. Rev.,* **102,** 684–704.

——, R. Hertwig, E. van den Broek, F. Fasolo, and K. V. Katsikopoulos, 2005: A 30% chance of rain tomorrow: How does the public understand probabilistic weather forecasts? *Risk Anal.,* **25,** 623–630.

Joslyn, S., and S. Savelli, 2010: Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteor. Appl.,* **17,** 180–195.

——, and J. LeClerc, 2012: Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol.,* **18,** 126–140, doi:10.1037/a0025185.

——, L. Nadav-Greenberg, and R. M. Nichols, 2009: Probability of precipitation: Assessment and enhancement of end-user understanding. *Bull. Amer. Meteor. Soc.,* **90,** 185–193.

Murphy, A. H., and R. L. Winkler, 1974: Credible interval temperature forecasting: Some experimental results. *Mon. Wea. Rev.,* **102,** 784–794.

——, S. Lichtenstein, B. Fischhoff, and R. L. Winkler, 1980: Misinterpretations of precipitation probability forecasts. *Bull. Amer. Meteor. Soc.,* **61,** 695–701.

National Research Council, 2006: Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts. National Research Council, 124 pp.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Roulston, M. S., G. E. Bolton, A. N. Kleit, and A. L. Sears-Collins, 2006: A laboratory study of the benefits of including uncertainty information in weather forecasts. *Wea. Forecasting,* **21,** 116–122.