# The Impact of Weather Forecast Inconsistency on User Trust

JESSICA N. BURGENO AND SUSAN L. JOSLYN

*Department of Psychology, University of Washington, Seattle, Washington*

(Manuscript received 19 June 2019, in final form 12 June 2020)

ABSTRACT

For high-impact weather events, forecasts often start days in advance. Forecasters believe that consistency among subsequent forecasts is important to user trust and can be reluctant to make changes when newer, potentially more accurate information becomes available. However, to date, there is little empirical evidence for an effect of inconsistency among weather forecasts on user trust, although the reduction in trust due to inaccuracy is well documented. The experimental studies reported here compared the effects of forecast inconsistency and inaccuracy on user trust. Participants made several school closure decisions based on snow accumulation forecasts for one and two days prior to the target event. Consistency and accuracy were varied systematically. Although inconsistency reduced user trust, the effect of the reduction due to inaccuracy was greater in most cases suggesting that it is inadvisable for forecasters to sacrifice accuracy in favor of consistency.

## 1. Introduction

Forecasts for major weather events often begin days in advance. The weather models upon which forecasts are based produce predictions that are updated periodically, generally changing and growing more accurate on average as lead times decrease (Lazo et al. 2009; Wilson and Giles 2013). However, when more recent model predictions contradict previous forecasts, meteorologists must decide whether or not to update the forecast they provide to the public. Sometimes they are reluctant to do so out of fear that inconsistency in subsequent forecasts (i.e., subsequent forecasts differ from the original forecast) will be confusing and negatively affect user trust.

Indeed, the maintenance of forecast consistency is considered important by many (Perry and Green 1982; Quarantelli 1984; Drabek 1999), including the National Oceanic and Atmospheric Administration (NOAA 2016). Moreover, evidence from outside of the weather domain suggests that consumers believe that consistency between two estimates from the same source is a signal of skill and should be maintained when reputation is at stake (Falk and Zimmermann 2017). In addition, people can detect trends in inconsistent forecasts that influence their expectations about future forecasts. There is evidence that people assume that trends have momentum such that an upward (4–6 in.) or downward (6–4 in.) revision will continue moving in the same direction into the future (Hohle and Teigen 2015; Erlandsson et al. 2018). Thus, it is clear that people are sensitive to changes in sequential forecasts from a single advisor.

There is also evidence that when people receive information about the same event from multiple sources, they prefer messages to be in agreement as opposed to conflicting, all else being equal (Smithson 1999). In addition, people have higher confidence in their own decisions when decisions are based on information from financial advisors who agree with one another as opposed to advisors who do not agree (Budescu et al. 2003). Nonetheless, when presented with conflicting financial advice from multiple advisors, participants' own estimates appear to be a simple average, suggesting that all of the advice was considered and weighted equally (Budescu and Yu 2007). It is important to note, however, that these are simultaneous forecasts from different sources that appear to "disagree" with one another (Løhre et al. 2019) rather than sequential forecasts from the same source. It could be argued that sequential inconsistency in a single source is fundamentally different in that the inconsistency arises from differences in prediction from that same source, reflecting more directly upon it.

Surprisingly, there is little experimental research that investigates the effect on trust of inconsistency in sequential forecasts in the weather domain per se. The one

*Corresponding author*: Jessica N. Burgeno, jburgeno@uw.edu

DOI: 10.1175/WCAS-D-19-0074.1

exception is an experiment that manipulated consistency in sequential thunderstorm and snow forecasts showing that increased forecast consistency led to greater trust in the forecasts (Losee and Joslyn 2018). In addition, there is evidence that when people receive multiple simultaneous weather warning messages from different sources, disagreement among them is confusing (Weyrich et al. 2019). There is also field evidence suggesting that conflicting evacuation orders for Hurricane Katrina led to lower perceived severity and failure to evacuate among African Americans (Elder et al. 2007). In sum, there is some preliminary evidence that consistency among weather forecasts may be important.

However, because accuracy (match between the forecast and the observed outcome) generally increases as lead times decrease, the choice to maintain consistency in sequential forecasts can be at the expense of accuracy. For example, in October 2016 historic and destructive winds were forecasted for Saturday, 15 October in western Washington State. The initial warning went out on Wednesday but by late Friday it was clear that the chance of an extreme event was decreasing. However, forecasters did not immediately downgrade the forecast they provided the public. By early Saturday morning the weather system decreased in size, moved further offshore, and although it was windy, extreme winds were not observed. As a result, the forecast was heavily criticized both locally and nationally as a gross exaggeration. This is just one of many similar examples. Although the intent was to preserve trust by providing consistent forecasts, meteorologists may have actually jeopardized public trust in future forecasts for major events and sacrificed accuracy in the process.

It is clear, in this example and in an abundance of experimental evidence, that forecast inaccuracy reduces trust. In one study, participants assigned a road salting task based on overnight low temperature forecasts reported significantly higher trust in low-error as compared to high-error forecasts and were more likely to take protective action (Joslyn and LeClerc 2012). In another study, participant investors rated higher competence and trustworthiness in accurate than inaccurate financial analysts and were more likely to purchase future reports from them (Kadous et al. 2009). In addition, mammography patients asked to imagine receiving an initial false positive breast cancer test result indicated diminished trust and greater likelihood of delaying future mammography relative to patients who imagined receiving accurate test results (Kahn and Luce 2003). Even preschoolers have been found to trust accurate more than inaccurate informants (Pasquini et al. 2007) and adjust their trust according to subsequent accuracy (Ronfard and Lane 2018).

Therefore, because consistency is widely advocated and the maintenance of consistency could be at the expense of accuracy in situations in which more recent information is more accurate, determining the relative impact of inconsistency on trust is critical. To that end, the three laboratory-based experiments reported below tested the following hypotheses: 1) inconsistency in sequential forecasts from the same source reduces user trust compared to consistent forecasts, 2) inaccuracy reduces trust compared to accurate forecasts, and 3) the reduction in trust due to inaccuracy is greater than the reduction due to inconsistency.

Participants were asked to take the role of a decision consultant responsible for advising schools whether or not to close due to snow based on sequential snow accumulation forecasts. Forecast consistency and accuracy were manipulated systematically to determine the impact on trust and closure decisions. Because of the tight control of extraneous variables required to examine these effects, no individual experiment was capable of fully addressing them. Therefore, these effects are addressed by the combined results of the three experiments.

## 2. Experiment 1

In experiment 1 forecast accuracy and consistency were manipulated in a computer-based task in which participants monitored sequences of weather forecasts in order to make school closure decisions. This allowed us to assess the impact of consistency and accuracy on trust ratings taken after learning the outcome on each trial.

### a. *Method*

#### 1) Participants

A total of 368 University of Washington psychology students (67% female, mean age = 19.1 years) participated for course credit and the opportunity to earn a cash bonus.

#### 2) Procedure

The experiment was programed in Excel Visual Basic and administered on standard desktop computers. Participants, tested in groups of about eight, first gave informed consent and provided their age and gender. Then they listened to, and read, instructions that described the task. Participants provided decision advice to schools regarding whether they should stay open or close due to an upcoming snowstorm. In reality, several factors are considered when making school closure decisions; however, in this simplified task, the decision was based on snow accumulation forecasts alone.

Participants were told to advise closing if they expected six or more inches of snow accumulation. Participants provided school closure advice over two hypothetical winter seasons, each lasting 12 weeks, for a total of 24 trials. Each week involved a different school district so that trials would be considered independent of one another.

To encourage engagement with the task, participants began with a virtual budget of 120 points and the goal of retaining as many points as possible by giving their best advice. Points could be spent at a rate of 2 per school closure recommendation to reflect the cost of makeup days. There was no cost for recommending that a school stay open; however, if participants recommended staying open and six or more inches of snow accumulation was observed, a 6-point penalty was deducted from their score to reflect the risk of travel in dangerous road conditions. Notice that, in the context of this task, as with weather situations in general, the cost of protection is less than the potential negative consequences of not protecting oneself. To further incentivize participants to put forth their best effort, cash was awarded for final balances at the rate of $1 for every 4 points over 72 (final balance) points. This payment threshold was chosen to discourage the simplistic and unrealistic strategy of recommending closure for every trial, which would result in a final balance of 72 points.

For each trial, participants were to base their school closure decision on two snow forecasts for Wednesday, a Monday forecast (two days prior), and a Tuesday forecast (one day prior). Because the effects of inconsistency on trust might be cumulative over trials it was blocked such that a sequence of six trials was either inconsistent or consistent and assigned to a particular, named forecast provider. There were four fictitious forecast providers—TruWeather, Weather Now, Weather Direct, and AccuCast—each of which, from the participants' perspective, was always either consistent or inconsistent. Before each new block, participants were notified of the new provider name. Forecast provider names were counterbalanced over the four forecast blocks; however, pretesting showed no significant difference in trust due to provider name alone.

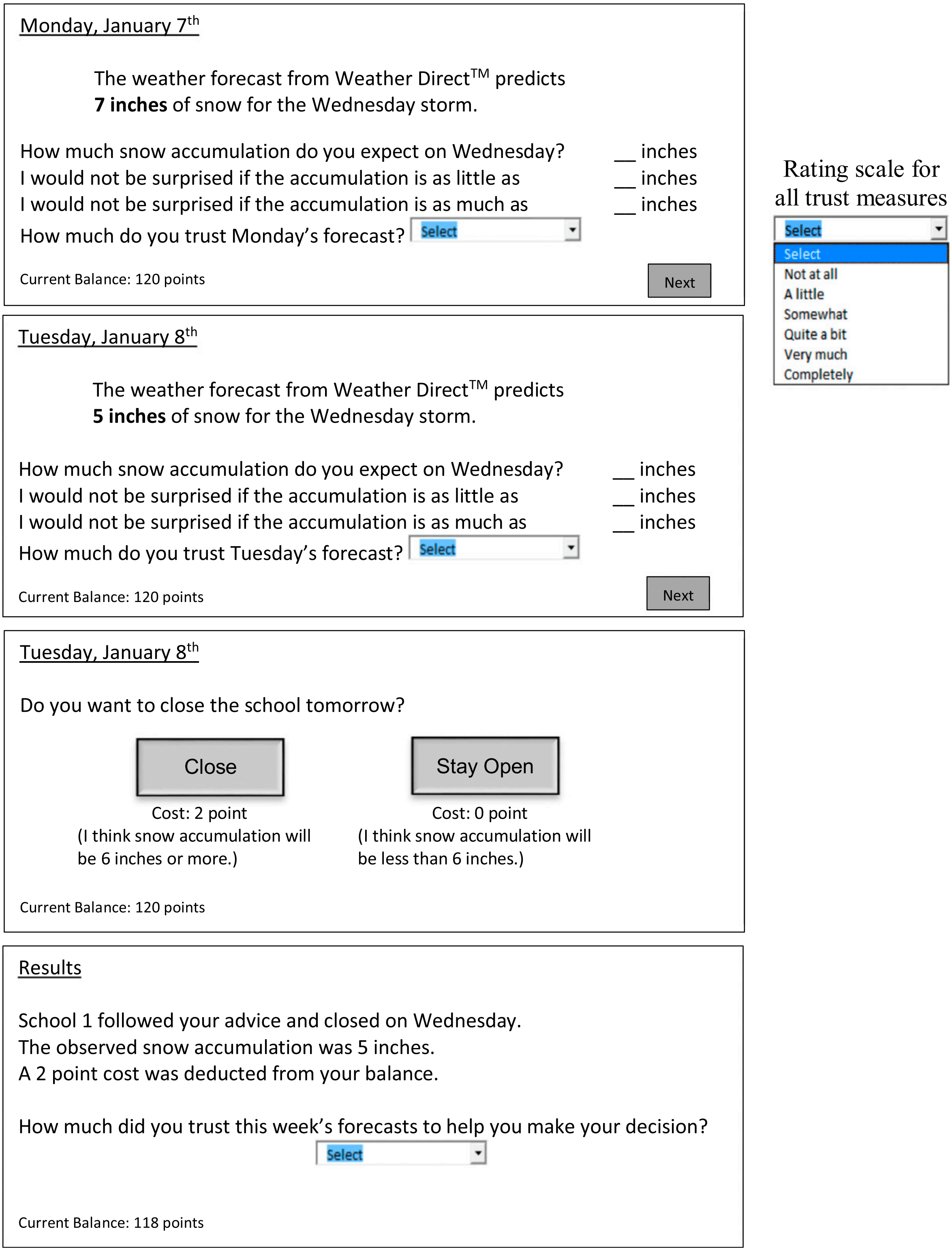
Each trial consisted of 4 screens: 1) Monday forecast for Wednesday, 2) Tuesday forecast for Wednesday, 3) Tuesday night school closure decision, and 4) a final screen in which participants were informed of the snow accumulation on Wednesday (see Fig. 1). The two snow accumulation forecasts for Wednesday were presented sequentially, centered on separate screens. The current date and day (Monday or Tuesday) appeared in the upper-left-hand corner of each screen in bold font. All dates were in the months of January, February, and March. Below each forecast (Monday and Tuesday), participants were asked to provide the number of inches of snow they expected for Wednesday, the least (minimum estimate) and greatest (maximum estimate) number of inches that they would not be surprised by, and to rate their trust in the forecast (''How much do you trust Monday's forecast?'') on a 6-point drop-down menu, from ''not at all'' to ''completely.'' Each participant's current point balance was shown in the lower-left-hand corner of every screen. When participants finished answering all four questions, they pressed a ''next'' button to advance to the next screen. At that point, they could not go back and change answers on the previous screens.

After the second forecast, participants were shown a decision screen. The current date and day (Tuesday) were shown in bold font in the upper-left-hand corner of the screen. A reminder of the Tuesday forecast for Wednesday was also provided in a box in the upper-right-hand corner of the decision screen. This was done to simulate actual situations in which decision-makers would likely have the Tuesday forecast available to them as they made the decision, whereas the Monday forecast would be a 24-hour-old memory subject to fading. Then, in the middle of the screen were two buttons labeled ''close'' or ''stay open.'' Text below each button reminded participants that closure meant ''I think snow accumulation will be 6 inches or more'' and that staying open meant ''I think snow accumulation will be less than 6 inches.'' Participants clicked on one of them to indicate their school closure decision.

After submitting their school closure decision, a fourth screen appeared stating that the school followed their advice. The observed snow accumulation was shown on the next line, and the resulting cost or penalty was indicated on the following line (unless neither occurred). Participants' point balance and, if applicable, the penalty incurred, was displayed in the lower-left corner of the screen. Here, participants once again rated their trust in the forecasts (''How much did you trust this week's forecasts to help you make your decision?'') using the same pull-down menu. Participants performed four practice trials before the test trials began.

### 3) Stimuli

There were four blocks of six trials. Each block had four experimental and two filler trials (explained below), for a total of 24 trials, 16 experimental, and 8 filler trials (see Table 1). The snow accumulation forecasts and observations consisted of realistic values for Washington State, where the experiment was conducted. Forecasted and observed values of snow accumulation in experimental trials ranged from 4 to 7 in. These values were



Monday, January 7th

The weather forecast from Weather Direct™ predicts
**7 inches** of snow for the Wednesday storm.

How much snow accumulation do you expect on Wednesday? __ inches
I would not be surprised if the accumulation is as little as __ inches
I would not be surprised if the accumulation is as much as __ inches
How much do you trust Monday's forecast? Select

Current Balance: 120 points

Tuesday, January 8th

The weather forecast from Weather Direct™ predicts
**5 inches** of snow for the Wednesday storm.

How much snow accumulation do you expect on Wednesday? __ inches
I would not be surprised if the accumulation is as little as __ inches
I would not be surprised if the accumulation is as much as __ inches
How much do you trust Tuesday's forecast? Select

Current Balance: 120 points

Tuesday, January 8th

Do you want to close the school tomorrow?

Cost: 2 point
(I think snow accumulation will be 6 inches or more.)

Cost: 0 point
(I think snow accumulation will be less than 6 inches.)

Current Balance: 120 points

Results

School 1 followed your advice and closed on Wednesday.
The observed snow accumulation was 5 inches.
A 2 point cost was deducted from your balance.

How much did you trust this week's forecasts to help you make your decision?
Select

Current Balance: 118 points

Rating scale for all trust measures

FIG. 1. Screens shown in a single trial in order from top to bottom.

TABLE 1. Inches of snow forecasted and observed by experiment and experimental trial type. Bold values highlight differences in forecast values across experiments. All other values were the same across all experiments. Participants were advised to close schools if they expected 6 in. or more of snow. Filler trials not included.

| | Accurate | | | Inaccurate | | | | |
|---|---|---|---|---|---|---|---|---|
| | First forecast | Second forecast | Observed outcome | First forecast | | | Second forecast | Observed outcome |
| Consistent | 4 | 4 | 4 | 4 | | | 4 | 6 |
| | 5 | 5 | 5 | 5 | | | 5 | 7 |
| | 6 | 6 | 6 | 6 | | | 6 | 4 |
| | 7 | 7 | 7 | 7 | | | 7 | 5 |
| | | | | E1 | E2 | E3 | | |
| Inconsistent | 4 | 6 | 6 | **4** | **3** | **7** | 5 | 7 |
| | 5 | 7 | 7 | **5** | **2** | **6** | 4 | 6 |
| | 6 | 4 | 4 | **6** | **9** | **5** | 7 | 5 |
| | 7 | 5 | 5 | **7** | **8** | **4** | 6 | 4 |

used because larger values are rare and might be distrusted for that reason, adding noise to the data.

There were four different types of experimental trials: accurate–consistent, accurate–inconsistent, inaccurate–consistent, and inaccurate–inconsistent. Half of the 4 experimental trials in each block were accurate and half were inaccurate. While there are other possible definitions for accuracy, for the purposes of this study accuracy was defined as an exact match between the Tuesday forecast and the accumulation observed on Wednesday. In all inaccurate experimental trials, there was a 2-in. difference between the second (Tuesday) forecast and the observed value. All inaccurate trials crossed the 6-in. closure threshold because an inaccuracy on the same side of the decision threshold could be considered accurate from the participants' perspective in that it would suggest the correct response. Trial order was randomized within a block.

Because the cost of failing to close when more than six inches of snow fell, was greater than the cost of closing, it was important to also hold constant the types of forecast errors. In half of inaccurate trials in each block the second forecast was at or above the 6-in. decision threshold and the observed accumulation was below it (false alarm). In the other half of inaccurate trials, the second forecast was below the 6-in. decision threshold and the observed accumulation was at or above the threshold (miss). Similarly, in half of accurate trials in each block both the second forecast and the observed accumulation values were below the 6-in. decision threshold (correct rejection). In the other half of accurate trials, the second forecast and the observed accumulation values were above the threshold (hit).

Half of experimental trials were consistent, and half were inconsistent. Out of concern that the effect of consistency may be small, each block of trials (attributed to a single forecast provider) contained exclusively consistent or inconsistent trials to allow for a buildup of trust or distrust over several trials. Consistency was defined as an exact match between the first and second forecasts. Inconsistent trials were inconsistent by 1.5 in. on average. Although ideal, it was not possible to match the magnitudes of inconsistency and inaccuracy for all trials while simultaneously controlling for forecasted and observed value ranges and ensuring that inaccurate trials crossed the 6-in. decision threshold. Therefore, in experiment 1 the inconsistencies in inaccurate trials were 1 in. while in accurate trials they were 2 in. (we return to this issue in the discussion). To control for any effects that might result from deducing trends over the two forecasts (Hohle and Teigen 2015, 2018; Maglio and Polman 2016), half of inconsistent trials had ascending forecasts (values increased from first to second forecast) and the other half had descending forecasts (values decreased from first to second forecast).

In an effort to obscure the regular patterns produced by controlling critical factors in the experimental trials, each block also included two filler trials. Filler trials were inaccurate by a 1-in. discrepancy between the second forecast and observed value and did not cross the 6-in. closure threshold. Filler trial values were lower (2 or 3 in.) or higher (8 or 9 in.) than values for experimental trials (4–7 in.).

There was also a forecast format manipulation. Half of participants received deterministic forecasts, and half received probabilistic forecasts. Deterministic forecasts were single-value forecasts implying an exact outcome (e.g., "4 inches of snow"). Probabilistic forecasts included both a single value forecast and a probability of six or more inches of snow accumulation (e.g., "4 inches of snow. . .however, there's a 30% chance of 6 or more inches of snow"). The probabilities for experimental forecasts ranged from 30% to 60%, in increments of 10, with a mean probability of 45%. In fact, 50% of trials at all probability levels resulted in an observed Wednesday snow accumulation at or above the 6-in. decision threshold. Therefore, the probabilistic forecasts were not reliable. Perhaps for that reason, we found no significant main effects or attenuating effects of forecast format. In all analyses reported below, the conditions were combined, and this manipulation will not be discussed further.

#### 4) Design

For the analyses reported below we used a 2 (accurate/inaccurate) by 2 (consistent/inconsistent) design. Accuracy and consistency were both within-group variables.

### b. Results

The primary hypotheses were that inconsistent forecasts would reduce trust compared to consistent forecasts, that inaccurate forecasts would reduce trust compared to accurate forecasts, and that the reduction in trust due to inaccuracy would be greater than that due to inconsistency. Where appropriate, Cohen's $d$ is provided to allow for effect size comparisons. Prior to conducting the main analyses, data for participants who did not understand the task or were not paying attention were omitted. To this end, we excluded the five participants who reported higher average minimum than maximum snow accumulation estimates leaving a total of 363 participants.

To compare the impact of consistency to that of accuracy directly, we first examined trust rated after the decision was made and the outcome of that decision was revealed. This set of analyses revealed that inconsistency did in fact significantly reduce trust, but not to the extent that inaccuracy did. The mean of trust ratings (taken *after* the outcome was revealed) was calculated for each trial type per participant. Then a 2 (accuracy: accurate, inaccurate) by 2 (consistency: consistent, inconsistent) repeated measures ANOVA was conducted on mean trust. Participants rated their trust in consistent forecasts [$M = 3.21$, standard deviation (SD) = 0.82] significantly higher than their trust in inconsistent forecasts ($M = 3.07$, SD = 0.88), independent of accuracy, $F(1, 361) = 17.35$, $p < 0.001$, Cohen's $d = 0.24$ (see Fig. 2). Likewise, participants rated their trust in accurate forecasts ($M = 3.35$, SD = 0.79) significantly higher than their trust in inaccurate forecasts ($M = 2.93$, SD = 0.94), independent of consistency, $F(1, 361) = 124.83$, $p < 0.001$, Cohen's $d = 0.79$. Notice that the magnitude of the effect of inaccuracy is substantially larger than that of inconsistency. Additionally, there was an unpredicted but significant interaction between consistency and accuracy showing a greater difference due to inconsistency when forecasts were accurate, $F(1, 361) = 8.78$, $p = 0.003$, Cohen's $d = 0.13$. Posthoc, Bonferroni corrected paired comparisons confirmed that consistency did not have a significant effect on trust when forecasts were inaccurate, $t(362) = 1.69$, $p = 0.092$, although the effect was significant when they were accurate, $t(362) = 4.94$, $p < 0.001$.

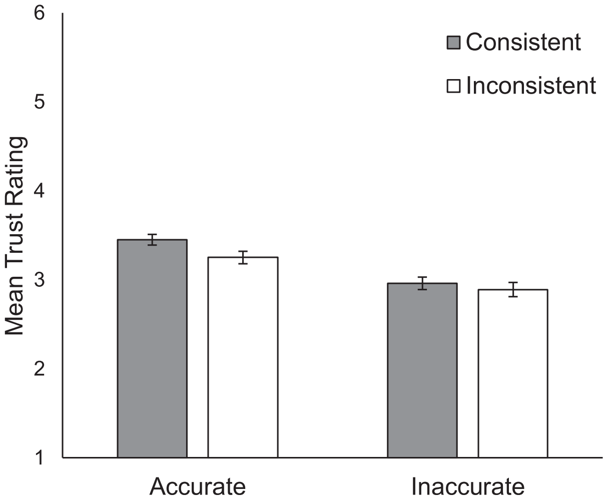


FIG. 2. Experiment 1 trust ratings by accuracy and consistency. Note: Error bars show the 95% confidence interval (CI).

The next analysis was conducted to compare the options forecasters might face in operational situations: to update a forecast (at the loss of consistency), or to maintain consistency in subsequent forecasts (potentially sacrificing accuracy). A paired samples $t$ test revealed significantly higher trust ratings in accurate-inconsistent forecasts ($M = 3.25$, SD = 0.91) than in inaccurate–consistent forecasts ($M = 2.96$, SD = 1.01), $t(362) = 5.72$, $p < 0.001$, Cohen's $d = 0.30$. This analysis is especially relevant considering that these trial types featured inaccuracies and inconsistencies with equal magnitudes (2 in.).

Although the effect of inconsistency on trust post outcome was small, it might be larger prior to learning about accuracy, when participants made their decisions. Surprisingly, however, a paired samples $t$ test revealed an even smaller effect of consistency on preoutcome trust. Although participants rated consistent forecasts significantly more trustworthy ($M = 3.19$, SD = 0.79) than inconsistent forecasts ($M = 3.05$, SD = 0.82), $t(362) = 4.51$, $p < 0.001$, the effect size was small, Cohen's $d = 0.17$.

We next investigated how participants incorporated the information in inconsistent forecasts into their own estimate of the outcome. Arguably the most recent forecast should be regarded as a replacement for the first forecast when it is different, as it is based on updated information (although this fact was not made explicit to participants). However, participants may put some weight on the first forecast or even weight both equally as has been seen in simultaneous predictions. A regression was conducted on participants' mean snow accumulation estimates with the first and second forecast values entered simultaneously as predictors. The second forecast clearly had a much stronger impact. A one unit increase in the second forecast predicted a 0.83 unit increase in snow accumulation estimates, $\beta = 0.47$, $p < 0.001$ while a one unit increase in the first forecast value predicted only a 0.11 unit increase, $\beta = 0.06$, $p < 0.001$. Note that the standardized beta coefficients indicate that the weighting of the second forecast was 7 times greater than that of the first. Overall, the two-predictor model accounted for 23% of the variance in snow accumulation estimates, $F(2, 2891) = 421.89$, $p < 0.001$, $R^2 = 0.23$.[1]

Thus, people seem to understand that when forecasts are inconsistent the second forecast is more relevant. They may also infer greater uncertainty when forecasts were inconsistent. To test this hypothesis uncertainty expectations were operationalized as the range of outcomes the participant would not find surprising. Ranges were calculated by subtracting participants' ''as little as'' estimates from their ''as much as'' estimates taken after the second forecast. A paired samples $t$ test on mean range revealed that participants estimated a significantly larger range for the target date when forecasts were inconsistent ($M = 3.53$, SD = 1.70) than when forecasts were consistent ($M = 3.35$, SD = 1.54), $t(362) = 3.97$, $p < 0.001$. It is important to note that consistent and inconsistent forecasts used the same forecast values the same number of times. In other words, we can be confident that this difference is due to consistency alone rather than the plausibility of values.

We blocked consistency to determine whether blocking increased its impact. If the effect of consistency were building over the course of a block in the hypothesized direction, the average trust in consistent forecasts would increase over trials within a block (positive correlation between trust and trial number) and the average trust in inconsistent forecasts would decrease over trials within a block (negative correlation between trust and trial number). With the exception of one block (the first block for participants who received an inconsistent block first, $r = -0.98$, $p = 0.006$), none of the correlations between trust and trial number reached significance. Although this is a sizeable correlation and blocking was included in subsequent experiments, no other significant effects due to blocking were found, so it will not be mentioned again.

### c. *Discussion*

These results suggest that with the forecast values used here, inconsistency negatively affects user trust but not to the extent that inaccuracy does. There was also evidence for an interaction between consistency and accuracy. Inconsistency mattered more when forecasts were accurate, suggesting that forecasters are ill advised to sacrifice accuracy for consistency. However, this interaction in particular may depend on the magnitude of inconsistencies in the stimuli used in experiment 1. Due to the constraints imposed by controlling for multiple variables simultaneously, inconsistencies were smaller when forecasts were inaccurate than when they were accurate. Second, all inaccuracies crossed the 6-in. decision threshold while inconsistencies crossed the threshold in only half of inconsistent trials (the accurate ones). This may have minimized the difference in trust between those consistent and inconsistent forecasts. In addition, it could account for the unpredicted interaction. Recall the inconsistency only mattered when forecasts were accurate where the 2-in., threshold-crossing inconsistencies occurred. Subsequent experiments were conducted to address these issues.

However, it is important to note that the crucial comparison that forecasters most likely face was unaffected by these issues. There was significantly greater trust in accurate inconsistent forecasts than in inaccurate consistent forecasts, in which the magnitudes of inconsistencies and inaccuracies were equal and both had values that crossed the 6-in. threshold. In addition, and somewhat surprisingly, the effect of inconsistency on preoutcome trust was small despite the fact that participants were as yet unaware of forecast accuracy. Moreover, blocking failed to enhance the impact of consistency in all but one of the eight blocks, suggesting that in most cases, any trust lost by inconsistency or gained by consistency does not extend to the next forecast. Taken together, this suggests that as far as user trust is concerned, forecasters may be better served by updating predictions when they believe that better information is available.

Moreover, these results suggest that participants glean some information from forecast inconsistency. They expect more uncertainty as evidenced by the wide range of outcomes anticipated. In addition, when forecasts are inconsistent, participants do not weight them equally. Instead the second (Tuesday) forecast had a much greater impact on participants snow total estimates than the first (Monday) forecast. This suggests that participants may have an intuitive understanding that the most recent forecast is likely to be more accurate.

## 3. Experiment 2

In experiment 2, the range of forecast values was expanded so that all inconsistencies and inaccuracies would differ by 2 in. In addition, experiment 2 tested

[1] We have provided information on the impact of forecast trends in the appendix for the interested reader.

the impact of the second forecast reminder that appeared on the decision screen in experiment 1. Although it was intended to simulate the greater availability of the current forecast compared to one viewed many hours previously in a real-world setting, the reminder might have had unintended effects on other variables. Therefore, in experiment 2 we also manipulated the reminder to test its impact. The computer-administered procedure was identical to that used for experiment 1.

### a. Method

#### 1) Participants

A total of 164 University of Washington psychology students (49.1% female, mean age = 19.71 years) who had not participated in the previous experiment, participated for course credit and the opportunity to earn a cash bonus.

#### 2) Stimuli

The stimuli were identical to experiment 1 with three exceptions (see Table 1). First, experiment 2 participants received only single-value, deterministic forecasts. Second, in experiment 2, the range of forecasted snow accumulation values for experimental trials was greater (2–9 instead of 4–7 in.) allowing for 2-in. inconsistencies throughout and making the magnitudes of inaccuracy and inconsistency equal for all trials. Nonetheless, mean forecast values remained equal across all trial types, and all other forecasted and observed snow accumulation values remained the same. Third, half of participants received reminders of the second forecast on the decision screen and half did not. However, there were no significant effects of forecast reminder therefore in all analyses reported below, the conditions were combined, and this manipulation will not be discussed further.

#### 3) Design

A 2 (accurate/inaccurate) by 2 (consistent/inconsistent) design was used. Accuracy and consistency were both within-group variables.

### b. Results

The same data omission criteria were used as in experiment 1. Two participants were omitted, leaving a total of 162 participants. Then, the main analyses were conducted using methods identical to experiment 1. Almost all of the results were replicated.

A 2 (accuracy: accurate, inaccurate) by 2 (consistency: consistent, inconsistent) ANOVA conducted on postoutcome trust indicated that (independent of accuracy) trust ratings for consistent forecasts ($M$ = 3.32, SD = 0.07) were significantly higher than for inconsistent forecasts ($M$ = 3.12, SD = 0.07), $F(1, 160) = 12.20$, $p$ = 0.001, Cohen's $d$ = 0.34 (see Fig. 3). In addition, (independent of consistency) trust ratings for accurate forecasts ($M$ = 3.42, SD = 0.06) were significantly higher than for inaccurate forecasts ($M$ = 3.03, SD = 0.07), $F(1, 160) = 54.58$, $p < 0.001$, Cohen's $d$ = 0.69. Again, the effect of inaccuracy was much greater than that of inconsistency. However, the accuracy by consistency interaction did not reach significance, $F(1, 160) = 2.21$, $p$ = 0.14, Cohen's $d$ = 0.09, although it trended in the expected direction. As with experiment 1, a paired samples $t$ test indicated that trust ratings for accurate inconsistent forecasts ($M$ = 3.29, SD = 0.07) were significantly higher than inaccurate consistent forecasts ($M$ = 3.10, SD = 0.08), $t(161) = 2.35$, $p$ = 0.020, Cohen's $d$ = 0.19.

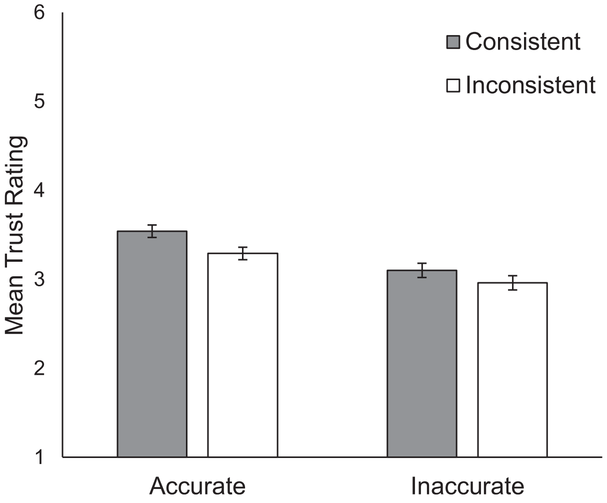


FIG. 3. Experiment 2 trust ratings by accuracy and consistency. Note: Error bars show the 95% CI.

In addition, preoutcome trust ratings were significantly higher for consistent ($M$ = 3.28, SD = 0.81) than inconsistent forecasts ($M$ = 3.09, SD = 0.74), $t(160)$ = 3.52, $p$ = 0.001, Cohen's $d$ = 0.24. As with experiment 1, the effect was smaller than the postoutcome effect of inconsistency (Cohen's $d$ = 0.34), and smaller than the effect of inaccuracy on postoutcome trust (Cohen's $d$ = 0.69).

To evaluate the impact of each forecast, when they were inconsistent, on participants own estimate of the outcome, a multiple regression analysis was conducted on snow accumulation estimates with first and second forecast values and forecast reminder entered simultaneously. Similar to experiment 1, the second forecast had a much bigger impact. A one unit increase in the second forecast value predicted a 0.85 unit increase in snow accumulation estimates, $\beta$ = 0.78, $p < 0.001$, while a one unit increase in the first forecast value predicted only a 0.04 unit increase in snow accumulation estimates, $\beta = 0.06, p < 0.001$. Overall, the two predictor model accounted for 67% of the variance in snow accumulation estimates, $F(2, 2589) = 2594.99$, $p < 0.001$, $R^2 = 0.67$.

Thus, inconsistent forecasts reduced trust and impacted snow accumulation expectations. We next examined whether inconsistent forecasts impacted school closure decisions (this analysis was precluded by the limited range of forecast values in experiment 1).[2] To increase the chance of detecting an effect due to consistency, we used an extreme groups design, including only the second forecast values of 4 in. (below threshold) and 7 in. (above threshold). This was done because decisions for values at or near the 6-in. closure threshold might be less clear-cut with respect to that threshold. Participants tend to anticipate slight error in the forecast that could be influenced by individual differences in risk tolerance (Joslyn and Savelli 2010). These slight differences in forecast interpretation would be less consequential to decisions for values further from the threshold allowing us to better detect the impact of consistency. Then, a 2 (consistency: consistent, inconsistent) by 2 (threshold orientation: below, above) repeated measures ANOVA was conducted on the mean percentage of school closure decisions. Indeed, participants closed significantly more often on *inconsistent* ($M = 0.61$, SD = 0.48) than on consistent forecasts ($M = 0.59$, SD = 0.49), $F(1, 160) = 9.83$, $p = 0.002$, Cohen's $d = 0.13$, suggesting a more cautious strategy when forecasts were inconsistent. Moreover, there was a significant consistency by second forecast interaction revealing that the effect occurred below (inconsistent $M = 0.23$, SD = 0.33; consistent $M = 0.09$, SD = 0.20) rather than above the threshold (inconsistent, $M = 0.91$, SD = 0.21; consistent, $M = 0.95$, SD = 0.16), $F(1, 160) = 27.60$, $p < 0.001$, Cohen's $d = 0.21$ (see Fig. 4).[3] It is important to note that the second forecast values and mean first forecast values were identical in consistent and inconsistent conditions ensuring that these effects were due to consistency alone. Not surprisingly, participants closed significantly more often above ($M = 0.93$, SD = 0.25) than below ($M = 0.16$, SD = 0.37) the threshold, $F(1, 160) = 1528.52$, $p < 0.001$, Cohen's $d = 3.74$.

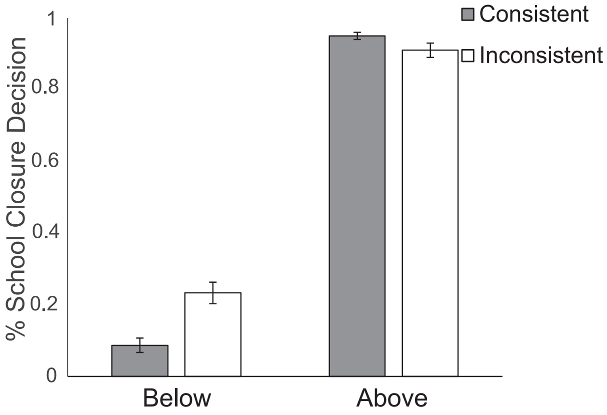


FIG. 4. Experiment 2 percent closed by threshold orientation and consistency.

The analysis examining whether inconsistency influenced participants uncertainty perceptions was omitted here because in order to address the confounds present in experiment 1, the first forecast values of some inconsistent conditions were different than some consistent conditions introducing a new confound that affected this analysis alone.

### c. *Discussion*

The negative effects of inconsistency and inaccuracy on user trust found in experiment 1 were replicated in experiment 2. Again, the magnitude of the effect of inaccuracy appears to be substantially larger than that of inconsistency. The consistency by accuracy interaction did not reach significance in experiment 2, although again there was a greater difference in inconsistency when forecasts were accurate. Moreover, in the crucial comparison between the options forecasters most often face, accurate but inconsistent forecasts were trusted significantly more than inaccurate consistent forecasts, as in experiment 1. In addition, as with experiment 1, without knowledge of accuracy, the effect of inconsistency on preoutcome trust was relatively small. Thus, the recommendation stands: as far as user trust is concerned, forecasters are better served updating their forecasts for the sake of accuracy.

As with experiment 1, participants' accumulation estimates were influenced more strongly by second forecast values than first forecast values. This suggests that, although they do not ignore the first forecast altogether, users understand that the second forecast should be regarded as a replacement for the first forecast and is likely to be more accurate.

[2] In experiment 1, all day 2 forecasts below the threshold (4 in.) were descending, and all day 2 forecasts above the threshold (7 in.) were ascending implying contradictory trends that might impact closure decisions.

[3] An analysis including all trials revealed a similar pattern with significantly more closures above than below the threshold, $F(1, 160) = 1053.46$, $p < 0.001$, and a significant consistency by threshold orientation interaction, $F(1, 160) = 31.38$, $p < 0.001$, suggesting that the effect of consistency was greater below than above the threshold. However, the effect of consistency failed to reach significance.

All in all, experiment 2 confirms the main results of experiment 1 suggesting that inconsistency reduces trust, although not to the degree of inaccuracy. In experiment 2, we found that inconsistency also impacted people's decisions causing them to be more cautious, advising school closure significantly more often when the second forecast was well below the decision threshold of 6 in.

However, one confound remained. Although all inaccuracies and inconsistencies were equal in magnitude in experiment 2, only half of inconsistencies crossed the 6-in. decision threshold (inaccurate–inconsistent) while all of the inaccuracies did. This could account for some of the differences observed here. In addition, a new confound was introduced in solving the magnitude problem. Although the mean forecast values were held constant, inaccurate inconsistent trials included first forecast values that were 1 and 2 in. higher and lower than the values of other trial types. We suspect that the impact of this change on trust was minimal because the smaller snow accumulation values would seem more plausible to western Washington residents, enhancing trust, while the larger values would seem less plausible making the combined effect essentially the same as the original values. Nonetheless, experiment 3 was conducted to correct for these confounds.

## 4. Experiment 3

Experiment 3 was conducted to determine whether the results of the previous experiments would hold, when all inaccuracies and inconsistencies were of equal magnitude (2 in.) *and* crossed the 6-in. decision threshold *and* forecast values were controlled. Although the second forecast reminder was manipulated once again, again we found no significant effects due to reminder and have combined these conditions in all analyses below. The procedure, design, and data summary methods were identical to experiment 2.

### a. Method

#### 1) Participants

A total of 160 University of Washington psychology students (50.6% female, mean age = 19.9 years), who had not participated in the previous experiments, participated for course credit and the opportunity to earn a cash bonus.

#### 2) Stimuli

The snow accumulation values in experiment 3 were identical to experiment 1 with one exception. In experiment 3, first forecast values in the four inaccurate inconsistent trials were allowed to match the outcome values so that 2-in. inconsistencies could cross the 6-in. decision threshold (e.g., first forecast: 4 in., second forecast: 6 in., observed: 4 in.). Although this is a somewhat unlikely (but not impossible) scenario, it is important to note that it occurred on a minority (17%) of trials and allowed us to resolve this important issue. Thus, the magnitudes of all inaccuracies and inconsistencies were equal (see Table 1) and all crossed the decision threshold.

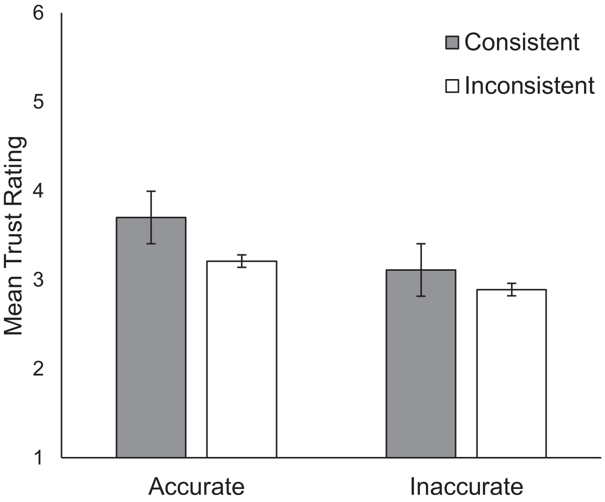


FIG. 5. Experiment 3 trust ratings by accuracy and consistency. Note: Error bars show the 95% CI.

### b. Results

The same data omission criteria were used as in experiments 1 and 2. Two participants were omitted, leaving a total of 158 participants. A 2 (accuracy: accurate, inaccurate) by 2 (consistency: consistent, inconsistent) ANOVA conducted on mean postoutcome trust revealed that consistent forecasts ($M = 3.40$, SD = 0.87) were rated significantly higher than inconsistent forecasts ($M = 3.05$, SD = 0.83) independent of accuracy, $F(1, 156) = 42.66, p < 0.001$, Cohen's $d = 0.61$. Accurate forecasts ($M = 3.45$, SD = 0.80) were rated significantly higher than inaccurate forecasts ($M = 3.00$, SD = 0.90), independent of consistency, $F(1, 156) = 66.68, p < 0.001$, Cohen's $d = 0.79$ (see Fig. 5). Notice that the effect of inaccuracy was again greater than that of inconsistency. The accuracy by consistency interaction was significant, as it had been in experiment 1, suggesting a greater difference in trust between consistent ($M = 3.70$, SD = 0.91) and inconsistent forecasts ($M = 3.21$, SD = 0.88) when the forecast was accurate than inaccurate (consistent, $M = 3.11$, SD = 1.08; inconsistent, $M = 2.89$, SD = 0.91), $F(1, 156) = 13.47$, $p < 0.001$, Cohen's $d = 0.22$. Bonferroni corrected paired comparisons revealed that although the difference was larger for accurate than

TABLE 2. Mean trust ratings by experiment and trial type. Bold text highlights means of pre- and postoutcome trust.

| Experiment | | Accurate | Inaccurate | Consistent | Inconsistent | Accurate consistent | Accurate inconsistent | Inaccurate consistent | Inaccurate inconsistent |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **Mean pre** | | | **3.19** | **3.05** | | | | |
| | Std dev | | | 0.79 | 0.82 | | | | |
| | **Mean post** | **3.35** | **2.93** | **3.21** | **3.07** | **3.45** | **3.25** | **2.96** | **2.90** |
| | Std dev | 0.79 | 0.904 | 0.82 | 0.88 | 0.86 | 0.91 | 1.01 | 1.01 |
| 2 | **Mean pre** | | | **3.28** | **3.09** | | | | |
| | Std dev | | | 0.81 | 0.74 | | | | |
| | **Mean post** | **3.42** | **3.03** | **3.32** | **3.12** | **3.54** | **3.29** | **3.10** | **2.96** |
| | Std dev | 0.78 | 0.93 | 0.85 | 0.88 | 0.86 | 0.90 | 1.05 | 1.01 |
| 3 | **Mean pre** | | | **3.33** | **3.01** | | | | |
| | Std dev | | | 0.77 | 0.81 | | | | |
| | **Mean post** | **3.45** | **3.00** | **3.41** | **3.05** | **3.70** | **3.21** | **3.11** | **2.89** |
| | Std dev | 0.80 | 0.90 | 0.87 | 0.83 | 0.91 | 0.88 | 1.08 | 0.91 |

inaccurate forecasts, it was significant for both (inaccurate, $t(157) = 3.27$, $p = 0.001$; accurate trials, $t(157) = 7.83$, $p < 0.001$). Although trust ratings for accurate inconsistent forecasts ($M = 3.21$, SD = 0.88) were higher than for inaccurate consistent forecasts ($M = 3.11$, SD = 1.08), unlike the first two experiments, the difference did not reach significance, $t(157) = 1.17$, $p = 0.244$, Cohen's $d = 0.10$. Again, the effect on preoutcome trust was small. A paired samples $t$ test revealed significantly higher preoutcome trust ratings for consistent ($M = 3.33$, SD = 0.77) than inconsistent trials ($M = 3.01$, SD = 0.81), $t(157) = 6.08$, $p < 0.001$, Cohen's $d = 0.39$. Thus, the main findings showing a greater impact of inaccuracy, as compared to inconsistency, on user trust were replicated here.

As with experiments 1 and 2, the impact of the first forecast on outcome estimates was small. A multiple regression on participants' mean snow accumulation estimates with first and second forecast values entered simultaneously as predictors revealed that the second forecast had a much bigger impact. A one unit increase in the second forecast value predicted a 0.79 unit increase in estimated snow accumulation, $\beta = 0.81$, $p < 0.001$, while a one unit increase in the first forecast value predicted a 0.13 unit increase in estimated snow accumulation, $\beta = 0.13$, $p < 0.001$. Overall, the two predictor model explained a significant proportion of the variance in snow accumulation estimates,[4] $F(2, 2524) = 2623.08$, $p < 0.001$, $R^2 = 0.68$.

### c. Discussion

Experiment 3 replicated nearly all of the effects on trust observed in the previous two experiments, with different stimuli designed to further address the confounds identified in the previous experiments. Again, there were significant negative effects of inconsistency and inaccuracy on user trust. Again, the magnitude of the effect of inaccuracy on trust was larger than that of inconsistency. Although the difference in effect sizes was reduced relative to experiments 1 and 2, the fact that it was observed in experiment 3 is particularly impressive. Recall that here, the first forecast matched the outcome in half of inaccurate experimental trials, which could have made those trials seem at least partially accurate to participants, reducing the effect of inaccuracy overall. However, matching the first forecast to the outcome was necessary to ensure that all inconsistent forecasts crossed the 6-in. threshold while maintaining control of the other extraneous variables.

Nonetheless, the consistency by accuracy interaction found in experiment 1 reemerged here suggesting that inconsistency matters more when forecasts are accurate than inaccurate, confirming that there is little benefit to consistency when accuracy is sacrificed. This could be because the effect of inaccuracy is so strong that it overwhelms any effect on trust of inconsistency. In addition, the effect of inconsistency on preoutcome trust remained relatively small, as in the previous two experiments (see Table 2). Taken together, these results contribute to the building evidence for the importance of accuracy over consistency in preserving user trust.

As in experiments 1 and 2, in experiment 3 participants' snow accumulation estimates were influenced more strongly by the second forecast values. In other words, participants appear to understand that the most recent forecast should be regarded as a replacement for the first.

## 5. General discussion

These three experiments, the first specifically designed to examine the relative effects of sequential forecast inconsistency and inaccuracy on trust, suggest that policies

[4] Neither school closure decision or expectation range analyses were possible in this experiment due to confounds in the stimuli mentioned in experiments 1 and 2.

TABLE 3. Postdecision trust analyses: Test statistics and effect sizes by experiment and effect. An asterisk (*) indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$.

| Experiment | | Accuracy | Consistency | Accuracy × consistency |
|---|---|---|---|---|
| 1 | *F* | 124.83*** | 17.35*** | 8.78** |
| | Cohen's *d* | 0.79 | 0.24 | 0.13 |
| 2 | *F* | 52.18*** | 12.59** | 1.94 |
| | Cohen's *d* | 0.69 | 0.34 | 0.09 |
| 3 | *F* | 66.65*** | 42.41*** | 13.72*** |
| | Cohen's *d* | 0.79 | 0.61 | 0.22 |

in favor of maintaining forecast consistency may be unwarranted in some cases. Because weather models tend to grow more accurate as lead times decrease, the artificial maintenance of forecast consistency can be at a cost to accuracy, which appears to be more important to user trust. In addition, inconsistent forecasts may provide users with important information about forecast uncertainty that can be applied to decision-making. Participants regard inconsistent forecasts as indicating greater uncertainty and are more likely to protect themselves.

To ensure that our effects were due to the primary independent variables, inaccuracy and inconsistency, we controlled for several extraneous variables including the forecast and observed values, whether forecast sequences ascended or descended, and error types. We also attempted to control the magnitudes of inaccuracies and inconsistencies, whether differences crossed the decision threshold, and the relationship of the forecasts to the outcome. Most but not all of these variables could be controlled in any given experiment. Nonetheless, the basic results held in all three experiments demonstrating their robustness and verifying that the effects reported here are due to inconsistency and inaccuracy per se, rather than to extraneous variables.

Granted the control of extraneous variables was done at some loss to ecological validity. However, this approach was necessary to fully understand the impact on trust of inconsistency and inaccuracy when all else is equal. It is also important to note that the extent of the inconsistencies and inaccuracies tested here was relatively small. It remains to be seen whether the pattern will hold for greater discrepancies. Nonetheless there is evidence that categorical inconsistencies (e.g., inconsistent forecast: a ''dusting of snow'' to ''several inches''; Losee and Joslyn 2018) yield a similar pattern of results suggesting that the effects may well be robust to different stimuli. Future studies should test stimuli with greater discrepancies and more naturalistic contexts to better understand how the results reported here interact with other factors. In addition, future studies should also test whether these results generalize beyond the weather domain (e.g., to climate change, medical, financial contexts, etc.), to different time horizons, and to decisions involving different consequences.

Across all three experiments inconsistency negatively impacted participants' trust in forecasts. Moreover, in each experiment the impact of inaccuracy was greater than that of inconsistency (see Table 3). However, our conclusions with respect to the relative size of the two effects (inaccuracy, inconsistency) is based primarily on experiment 3. Recall that in experiments 1 and 2 although all inaccuracies crossed the decision threshold, only half of inconsistencies did so. Importantly, this confound was eliminated in experiment 3, where the effect of inaccuracy continued to exceed that of inconsistency. This is particularly impressive because in order to eliminate the threshold-crossing confound, another confound was created: In half of the inaccurate forecasts the first forecast matched the outcome, potentially making them at least partially accurate in the eyes of participants.

In experiments 1 and 3, the significant interaction between accuracy and consistency suggested that consistency matters mainly when forecasts are accurate. However, this conclusion as well, rests primarily on experiment 3. In experiment 1 the inconsistent trials in the accurate condition crossed the decision threshold while those in the inaccurate condition did not, potentially reducing the impact. Importantly, the interaction was also observed in experiment 3 where this confound was eliminated, suggesting that indeed inconsistency is more important for accurate than for inaccurate forecasts. This may be because the negative impact of inaccuracy on trust is so powerful that it overwhelms the impact of inconsistency. Indeed, if consistency effects trust because it is regarded as a signal of skill, as previous work has suggested (Falk and Zimmermann 2017), inaccuracy may negate that impression. Taken together, these results suggest that any gain in trust from consistency may well be lost if the forecast turns out to be inaccurate.

We first tested postoutcome trust in order to compare inconsistency directly to the impact of inaccuracy.

However, from a practical standpoint, participants' trust in the forecast prior to learning the outcome may be more important to the choice they make. This was reflected in the preoutcome trust rating. Here too, in all three experiments, consistency had only a small effect, smaller than the effect on postoutcome trust and much smaller than the effect on trust due to inaccuracy. This contradicts the intuition that the diagnostic relevance of consistency (Falk and Zimmermann 2017) should be *greater* in the absence of accuracy information. One possible explanation for the smaller impact of inconsistency pre- than postoutcome, is that in the inconsistent forecast pairs, when the second forecast was accurate (by the definition used here) the first forecast was inaccurate. Therefore, inconsistent forecast pairs were less accurate overall and perhaps less trustworthy for that reason.

Contrary to our intuition, there was little evidence that the effect of consistency built over trials (blocking). If the effect of consistency were building over the course of a block, the average trust in consistent forecasts would increase and the average trust in inconsistent forecasts would decrease over the block. In only one of the 24 blocks was there a significant correlation between inconsistency and trial number. Nevertheless, the effect of consistency was significant in all three experiments, suggesting that it is not dependent on blocking. This may also suggest that, to the degree that trust was affected by forecast consistency, participants regarded it as a characteristic of the forecast rather than the forecast provider.

It is also clear that inaccuracy significantly decreases trust. This effect was found across all three experiments, regardless of the variation in stimuli. In addition, inaccuracy may have impacted trust in subsequent forecasts. Notice that even forecasts that were both consistent and accurate were not rated fully trustworthy, perhaps because of inaccurate trials preceding them. Indeed, the negative effects of inaccuracy on trust have been shown to endure long after accuracy improves (Joslyn and LeClerc 2012). In addition, it is important to realize that the relative impact of inaccuracy may be even greater in natural settings where this variable is not held constant. Here accuracy was exactly 50% for both consistent and inconsistent forecasts. In a natural setting the more recent forecast would likely be more accurate. Therefore, forecasts held artificially consistent by the forecaster would likely be less accurate on average than forecasts that were updated (inconsistent), further reducing trust.

Moreover, when forecasters artificially maintain consistency, they may be depriving users of potentially important decision-relevant information. Experiments 1 and 3 demonstrated that people expected a larger range of outcomes when forecasts were inconsistent relative to when they were consistent, suggesting that inconsistency may be taken as an indication of uncertainty. In experiment 2, people made more cautious decisions when forecasts were inconsistent, especially when the forecast predicted low snow totals, below the decision threshold. Notice that this result contradicts the survey evidence showing that inconsistent messaging leads to failure to take protective action (Elder et al. 2007). This difference may be due to the multiple other factors influencing decisions in natural settings or to a different operationalization of inconsistency. In the experiments reported here, inconsistency referred to differences in weather outcomes per se (snow accumulation) rather than advice about what to do (evacuate). Although this difference seems subtle, it is possible that inconsistency in advice is less well tolerated.

Indeed, we are not claiming that consistency in general is ill advised when communicating information to lay audiences. Consistency in terminology and presentation format make it easier for users to access and interpret similar information. The advantages of these forms of consistency are well documented (Oonk et al. 2001). It may be that consistency in advice is also important. This is a question that future experimental research should pursue. Moreover, due the small but replicable effect of inconsistency on trust reported here, if for some reason, accuracy is not an issue, maintaining consistency in forecast values can be beneficial. However, because prioritizing consistency often means deprioritizing accuracy, the costs of maintaining forecast consistency could easily outweigh the benefits. In sum, the experiments reported here suggest that not only is inconsistency in forecast values less deleterious to trust than inaccuracy, but it may also provide the user with important information.

In addition to the impact of inconsistency and inaccuracy, we were interested in how people integrate information from differing forecasts. In all three experiments, the weighting of the second forecast was at least 7 times greater than the earlier forecast. This suggests that participants understand that more recent information is likely to be more accurate and therefore they emphasize the second forecast in their own estimate. This could be due to extraexperimental experience with real weather forecasts about which people have many, often valid intuitions (Morss et al. 2008; Joslyn and Savelli 2010; Savelli and Joslyn 2012). However, within the experimental setting, our forecast stimuli were realistic in that sense. Second forecasts tended to be more accurate (50% accurate) than first forecasts (25% accurate). Participants might have learned (explicitly or implicitly) to discount first forecasts as "mostly wrong." Thus,

TABLE A1. Mean and SD snow accumulation estimates by consistent and inconsistent conditions. Inconsistent conditions are broken down by ascending and descending categories.

| Snow estimate | Expt 1 ($N$ = 363) | | Expt 3 ($N$ = 158) | |
|---|---|---|---|---|
| Consistent | $M$ = 5.42, SD = 1.21<br>Trials = 8 | | $M$ = 5.45, SD = 1.22<br>Trials = 8 | |
| Inconsistent | Ascending<br>$M$ = 6.05<br>SD = 1.02<br>Trials = 4 | Descending<br>$M$ = 4.96<br>SD = 2.49<br>Trials = 4 | Ascending<br>$M$ = 6.14<br>SD = 0.78<br>Trials = 4 | Descending<br>$M$ = 4.78<br>SD = 0.87<br>Trials = 4 |

unlike simultaneous predictions from separate sources that tend to be weighted equally (Budescu and Yu 2007), the most recent forecast is much more heavily weighted for sequential forecasts from the same source, suggesting that information integration strategies may differ depending on the temporal relationship (simultaneous vs sequential) or source (single vs multiple sources) of the decision information. Nonetheless, the differential weighting observed in the experiments reported here may explain the smaller effect of inconsistency on trust relative to inaccuracy. Perhaps, because people understand that the more recent forecast is likely to be more accurate, the difference across forecasts matters less to them.

As such, these results have implications for a broad range of domains that involve sequential predictions. They suggest that although inconsistency in information can have a negative effect on trust, providers of such information should not artificially preserve consistency at a potential loss to accuracy. Most people likely understand that forecasts change and grow more accurate as more information becomes available. Indeed, our participants depended far more heavily on the second than on the first forecast. Thus, updating predictions, even at the expense of consistency, can preserve trust in the information source as well as provide users with higher quality decision-relevant information.

*Acknowledgments.* This research was supported by the National Science Foundation under Grant 1559126. Data for all experiments can be found at https://osf.io/rzbcw.

## APPENDIX

### The Impact of Forecast Trends

Although the experiments reported here were not specifically designed to test this question, we provide analyses of the effect of ascending and descending trends in forecasts (all in the inconsistent condition) on snow accumulation estimates and school closure decisions for experiments 1 and 3 where forecast values were not confounded with these categories.

Indeed, snow accumulation estimates for ascending forecasts were significantly larger than for descending forecasts in both experiment 1, $t(362) = 13.86$, $p < 0.001$, and 3, $t(157) = 21.57$, $p < 0.001$ (see Table A1). However, very few estimates in each category continued the trend. In experiment 1, only 10% of estimates for ascending trials were larger than the second forecast. Only 12% of estimates for descending trials were smaller than the second forecast. Likewise, in experiment 3 only, 4% of estimates for ascending trials were larger than the second forecast and only 8% of estimates in descending trial were smaller than the second forecast. Thus, while a few of these estimates may be due to anticipating trends, it is likely that most are better explained by greater weighting on the second over the first forecast, due to its recency.

Similarly, participants closed significantly more often when forecasts were ascending compared to when they were descending in both experiment 1, $t(362) = 24.43$, $p < 0.001$, and 3, $t(157) = 13.46$, $p < 0.001$ (see

TABLE A2. Mean and SD school closure decisions by consistent and inconsistent conditions. Inconsistent conditions are broken down by ascending and descending categories.

| Percent Closed | Expt 1 ($N$ = 363) | | Expt 3 ($N$ = 158) | |
|---|---|---|---|---|
| Consistent | $M$ = 0.57, SD = 0.50<br>Trials = 8 | | $M$ = 0.57, SD = 0.50;<br>Trials = 8 | |
| Inconsistent | Ascending<br>$M$ = 0.78<br>SD = 0.42<br>Trials = 4 | Descending<br>$M$ = 0.40<br>SD = 0.49<br>Trials = 4 | Ascending<br>$M$ = 0.78<br>SD = 0.50<br>Trials = 4 | Descending<br>$M$ = 0.42<br>SD = 0.41<br>Trials = 4 |

TABLE A3. Experiment 1 ($N = 363$) mean and SD trust ratings by accuracy and consistency. Inconsistent/inaccurate conditions are broken down by trend verified and trend contradicted categories.

| Trust | Accurate | Inaccurate | |
|---|---|---|---|
| Consistent | $M = 3.45$<br>SD = 1.22<br>Min = 0<br>Max = 6<br>Trials = 4 | $M = 2.96$<br>SD = 1.34<br>Min = 0<br>Max = 6<br>Trials = 4 | |
| Inconsistent | | Verified | Contradicted |
| | $M = 3.25$<br>SD = 1.19<br>Min = 0<br>Max = 6<br>Trials = 4 | $M = 2.93$<br>SD = 1.29<br>Min = 0<br>Max = 6<br>Trials = 2 | $M = 2.86$<br>SD = 1.35<br>Min = 0<br>Max = 6<br>Trials = 2 |

Table A2). However, based on the analysis above, in most cases this was likely due, not to the trend per se, but rather to the systematically higher values in the second forecast (see Table 2) in the ascending as compared to descending pairs, which was weighted more heavily by participants.

We also analyzed of the effect of verified and contradicted trends in forecasts (all in inconsistent/inaccurate condition) on trust ratings for experiment 1, the only experiment where outcome expectations (based on forecast trend) were both verified and contradicted. However, the difference in trust between trend validated (outcome continued the trend) and trend contradicted (outcome contradicted the trend) trials failed to reach significance, $t(362) = 1.23$, $p = 0.22$ (see Table A3), suggesting that our primary findings are not explained by the trend effect.

## REFERENCES

Budescu, D. V., and H. T. Yu, 2007: Aggregation of opinions based on correlated cues and advisors. *J. Behav. Decis. Making*, **20**, 153–177, https://doi.org/10.1002/bdm.547.

——, A. K. Rantilla, H. T. Yu, and T. M. Karelitz, 2003: The effects of asymmetry among advisors on the aggregation of their opinions. *Organ. Behav. Hum. Decis. Process.*, **90**, 178–194, https://doi.org/10.1016/S0749-5978(02)00516-2.

Drabek, T. E., 1999: Understanding disaster warning responses. *Soc. Sci. J.*, **36**, 515–523, https://doi.org/10.1016/S0362-3319(99)00021-X.

Elder, K., S. Xirasagar, N. Miller, S. A. Bowen, S. Glover, and C. Piper, 2007: African Americans' decisions not to evacuate New Orleans before Hurricane Katrina: A qualitative study. *Amer. J. Public Health*, **97** (Suppl. 1), S124–S129, https://doi.org/10.2105/AJPH.2006.100867.

Erlandsson, A., S. M. Hohle, E. Løhre, and D. Västfjäll, 2018: The rise and fall of scary numbers: The effect of perceived trends on future estimates, severity ratings, and help-allocations in a cancer context. *J. Appl. Soc. Psychol.*, **48**, 618–633, https://doi.org/10.1111/jasp.12552.

Falk, A., and F. Zimmermann, 2017: Consistency as a signal of skills. *Manage. Sci.*, **63**, 2197–2210, https://doi.org/10.1287/mnsc.2016.2459.

Hohle, S. M., and K. H. Teigen, 2015: Forecasting forecasts: The trend effect. *Judgm. Decis. Making*, **10**, 416–428.

——, and ——, 2018: When probabilities change: Perceptions and implications of trends in uncertain climate forecasts. *J. Risk Res.*, **22**, 555–569, https://doi.org/10.1080/13669877.2018.1459801.

Joslyn, S. L., and S. Savelli, 2010: Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteor. Appl.*, **17**, 180–195, https://doi.org/10.1002/met.190.

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

Kadous, K., M. Mercer, and J. Thayer, 2009: Is there safety in numbers? The effects of forecast accuracy and forecast boldness on financial analysts' credibility with investors. *Contemp. Account. Res.*, **26**, 933–968, https://doi.org/10.1506/car.26.3.12.

Kahn, B. E., and M. F. Luce, 2003: Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Mark. Sci.*, **22**, 393–410, https://doi.org/10.1287/mksc.22.3.393.17737.

Lazo, J. K., R. E. Morss, and J. L. Demuth, 2009: 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bull. Amer. Meteor. Soc.*, **90**, 785–798, https://doi.org/10.1175/2008BAMS2604.1.

Løhre, E., A. Sobkow, S. M. Hohle, and K. H. Teigen, 2019: Framing experts' (dis)agreements about uncertain environmental events. *J. Behav. Decis. Making*, **32**, 564–578, https://doi.org/10.1002/bdm.2132.

Losee, J. E., and S. Joslyn, 2018: The need to trust: How features of the forecasted weather influence forecast trust. *Int. J. Disaster Risk Reduct.*, **30**, 95–104, https://doi.org/10.1016/j.ijdrr.2018.02.032.

Maglio, S. J., and E. Polman, 2016: Revising probability estimates: Why increasing likelihood means increasing impact. *J. Pers. Soc. Psychol.*, **111**, 141–158, https://doi.org/10.1037/pspa0000058.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2008: Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

NOAA, 2016: Risk communication and behavior: Best practices and research findings. NOAA Tech. Rep., 60 pp., https://www.performance.noaa.gov/wp-content/uploads/Risk-Communication-and-Behavior-Best-Practices-and-Research-Findings-July-2016.pdf.

Oonk, H. M., H. S. Smallman, and R. A. Moore, 2001: Evaluating the usage, utility and usability of web-technologies to facilitate knowledge sharing. *Proc. 2001 Command and Control Research and Technology Symp.*, Annapolis, MD, U.S. Department of Defense, http://www.dodccrp.org/events/6th_ICCRTS/Tracks/Papers/Track4/072_tr4.pdf.

Pasquini, E. S., K. H. Corriveau, M. Koenig, and P. L. Harris, 2007: Preschoolers monitor the relative accuracy of informants. *Dev. Psychol.*, **43**, 1216–1226, https://doi.org/10.1037/0012-1649.43.5.1216.

Perry, R. W., and M. R. Green, 1982: The role of ethnicity in the emergency decision-making process. *Sociol. Inq.*, **52**, 306–334, https://doi.org/10.1111/j.1475-682X.1982.tb01257.x.

Quarantelli, E. L., 1984: Perceptions and reactions to emergency warnings of sudden hazards. *Ekistics*, **51**, 511–515.

Ronfard, S., and J. D. Lane, 2018: Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. *Child Dev.*, **89**, 414–429, https://doi.org/10.1111/cdev.12720.

Savelli, S., and S. Joslyn, 2012: Boater safety: Communicating weather forecast information to high-stakes end users. *Wea. Climate Soc.*, **4**, 7–19, https://doi.org/10.1175/WCAS-D-11-00025.1.

Smithson, M., 1999: Conflict aversion: Preference for ambiguity vs conflict in sources and evidence. *Organ. Behav. Hum. Decis. Process.*, **79**, 179–198, https://doi.org/10.1006/obhd.1999.2844.

Weyrich, P., A. Scolobig, and A. Patt, 2019: Dealing with inconsistent weather warnings: Effects on warning quality and intended actions. *Meteor. Appl.*, **26**, 569–583, https://doi.org/10.1002/MET.1785.

Wilson, L. J., and A. Giles, 2013: A new index for the verification of accuracy and timeliness of weather warnings. *Meteor. Appl.*, **20**, 206–216, https://doi.org/10.1002/met.1404.