The Effect of Forecast Inconsistency and Explicit Uncertainty Estimates on Trust and Decision-Making

Jessica Noel Burgeno

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading committee:

Susan Joslyn, Chair

Andrea Stocco

Ann Bostrom

Program Authorized to Offer Degree:

Psychology

University of Washington

Abstract

The Effect of Forecast Inconsistency and Explicit Uncertainty Estimates on Trust and Decision-Making

Jessica Noel Burgeno

Chair of the Supervisory Committee:

Susan Joslyn

Department of Psychology

The goal of this dissertation was to evaluate whether inconsistency is a substantial threat to trust in sequential predictions from a single advisor, and to test whether providing explicit uncertainty estimates can preserve trust in the context of inconsistency. Four experiments were conducted in which participants decided, according to a sequence of snow accumulation forecasts, whether or not they should close schools due to an impending snowstorm. Since prioritizing consistency in weather forecasts can be at a tradeoff to forecast accuracy, Experiments 1-3 used highly controlled forecast stimuli to directly compare the relative effects of inaccuracy and inconsistency on trust. In Experiment 4, findings from the highly controlled experiments were retested and extended with realistic forecast stimuli varying naturally in terms of inaccuracy and inconsistency. In addition, half of participants were randomly assigned to receive reliable uncertainty estimates for a threshold level of snow with the deterministic forecast. Across all

four experiments inconsistency was less detrimental to trust than inaccuracy and appeared to confer decision-making benefits. Access to reliable uncertainty estimates amplified these benefits.

Acknowledgments

This dissertation work was made possible by funding from the National Science Foundation.

I am eternally grateful for the support of my advisor, Dr. Susan Joslyn, throughout this dissertation work. Thank you for nurturing my enthusiasm and development as an independent researcher, and for reinforcing my commitment to use my skills for social good. It has been a privilege to work alongside you and on such interesting and topical questions.

This dissertation also benefited from the attention and feedback of my research assistants and lab mates. I could not have completed this dissertation without the research assistants who collected most of the data. Many thanks, Autumn Spriggle, Shuya Dai, Shihan Xu, Chen Su, Mengying Xu, Justin Takeuchi, Keiko Shannon, Eva Yin, Ann Karmol, and Brandy Steed! Also thanks to my lab mates Dr. Sonia Savelli, Dr. Raoni Demnitz, Gala Gulacsik, Chao Qin, Jee Hoon Han, Dr. Meg Grounds, and Dr. Jared LeClerc, for sharing your work, wisdom, and perspectives!

I would also like to thank my committee members, Dr. Andrea Stocco and Dr. Ann Bostrom. Your thoughtful questions and comments refreshed and expanded my thoughts on my dissertation work. Thank you for your stimulating contributions!

I am forever indebted to my parents, Rick and Deb, who made this path less difficult to choose. Thank you for continuously lowering barriers and simultaneously instilling in me the value of hard work and perseverance. Also, many thanks to all my family for their support and confidence in me.

Finally, I am concluding this journey with the aid of my partner, Bernie Jara. Many thanks for maintaining my morale with love and light, and for taking on your many support roles

with grace. Thank you for bringing me closer to balance, and for making everything an adventure.

Table of Contents

Chapter 1. INTRODUCTION

High stakes decisions are frequently made under tremendous uncertainty about the future outcome or one's best course of action (e.g., to take protective action or not). These are the situations in which reliable risk communication can be critical, making it imperative that decision-relevant information is communicated in a way that supports understanding and trust. For example, as the Covid-19 pandemic unfolded in early 2020, there were extensive uncertainties around transmission of the virus punctuated by supply-chain issues. The World Health Organization's (WHO) initial mask message on January 29, 2020 suggested that face masks offered limited utility as a preventative measure against the spread of Coronavirus for members of the general public (WHO, 2021). Even in light of emerging evidence supporting pre-symptomatic and asymptomatic transmission and suggesting that masks may limit the spread of the virus, in a subsequent message on April 6th, 2020 the WHO maintained that they did not advise face masks for members of the general public, only healthcare workers and people known to be infected with the virus (WHO, 2021). The intent was at least partially to reserve masks for health care workers (Jingnan, 2020). However, by avoiding the slightly more complicated truth, that masks are effective against the spread of Coronavirus but that members of the public should refrain from using masks to save them for health care workers, they may have jeopardized their mission to support public health. In addition, by downplaying the utility of face masks for members of the general public, the WHO exposed themselves to later criticisms of inconsistency and inaccuracy.

In line with this example, concern over burdening laypeople with too much information has historically led to simplified communications. This can result in the omission of information, for instance about the extent of uncertainty in the situation (NRC, 2006), which

might help people make better decisions. Despite these concerns, the present research suggests that people are capable of understanding and making effective use of fairly complex scientific information, and that they trust it to a greater degree than simplified messages.

Like the Covid-19 example above, the dissertation research presented here focusses on information from a single source that may change over time. However, my dissertation studies were conducted in the context of the weather domain in which the appropriate provision of information is similarly a concern with potentially high-stakes consequences. Forecasts for major weather events often begin days in advance and change over time as more reliable information becomes available. Weather forecasts are based on numerical models which take as input current conditions and apply the principles of atmospheric physics to generate forecasts. In general, more recent model outputs tend to be more accurate (that is, the most recent forecast is closer to the observed outcome more often) than previous model runs because updated models are based on more recent information (Lazo et al. 2009; Wilson & Giles, 2013). However, forecasters are often hesitant to update forecasts out of fear that inconsistency (a mismatch) in subsequent forecast values for the same date will be confusing and negatively impact user trust in forecasts. This preference for the maintenance of consistency creates a potential tradeoff in the sense that if more recent forecasts are on average more accurate, artificially maintaining consistency by not updating can be at a loss to accuracy.

Consider a real-world example of the tradeoff between forecast accuracy and consistency. In 2016 hazardous and extraordinary wind speeds were forecasted in western Washington state for Saturday October 15th. In the days following the initial Wednesday forecast, the weather model updated indicating that the probability of substantial winds had reduced considerably. Yet, forecasters opted not to downgrade the public-facing forecast. When that Saturday morning

arrived, the storm had reduced in magnitude and migrated further seaward. Although it was windy, extraordinary winds were not observed. As a result, the inflated forecast was criticized widely, both locally and nationally. While the intention was to preserve trust by maintaining forecast consistency forecasters might have sown distrust in future forecasts for extreme weather events by failing to prioritize forecast accuracy.

Yet, it is not entirely surprising that forecasters are motivated to avoid inconsistencies. One only needs to recall the backlash that arose from the inconsistent messaging about face coverings as a tool to prevent the spread of Coronavirus in 2020, to realize the potential consequences for trust of changing advice over time. Afterall, maintaining forecast consistency is considered best practice by distinguished institutions, such as the National Oceanic and Atmospheric Administration (NOAA, 2016).

1.1 Trust

One of the biggest concerns about inconsistency in messaging is that it will negatively impact trust. There are several different kinds of trust that fall under at least two major categories, trust in motives and trust in competence (Twyman et al. 2008). The research reported here focusses on the latter, using a construct like that described by Earle (2010), to reflect trust based on forecasters' past performance, abilities, or knowledge. Note also that the focus of the present research is on the impact of inconsistency on trust in information from a single source as opposed to inconsistency in information from multiple sources. Whereas inconsistencies across different sources may reflect more on the situation (e.g., the extent of uncertainty involved), inconsistencies in information from a single source may reflect more directly on that source.

1.1.1 Inconsistency

There are different kinds of inconsistency. While consistency in formatting and terminology can make it easier for users to access and interpret information (Oonk et al., 2001),

consistency of facts may pose a problem if in fact it gives rise to greater inaccuracy. The present research demonstrates that convincingly. Despite widespread belief that inconsistency in facts is deleterious to trust, prior to this dissertation work little experimental research had tested the impact of inconsistency on trust. One exception is a study which manipulated forecast consistency for thunderstorm and snow forecasts discovered that forecast inconsistency decreased trust, as expected (Losee & Joslyn, 2018). However, in an experiment which manipulated the consistency of simultaneous snow forecasts from 2 distinct forecasters 1-day prior to a predicted storm discovered that inaccuracy substantially reduced user trust while inconsistency did not (Su, Burgeno, & Joslyn, 2021). Together these studies suggest that inconsistency may be more problematic for trust in information from a single source than from multiple sources. One of the goals of this dissertation is to systematically test the impact of inconsistency on trust in information from a single source to establish whether the impact of inconsistency on trust is indeed negative.

1.1.2 Inaccuracy

In contrast to the impact of inconsistency on trust, the question of whether inaccuracy reduces trust is well explored and extensive evidence indicates that indeed, forecast inaccuracy reduces trust. For example, in a study which tasked participants with making road salting decisions based on overnight low temperature forecasts, participants reported significantly higher trust and took protective action more often when faced with low-error versus high-error forecasts (Joslyn & LeClerc, 2012). Likewise, in a study which tasked participants with making investment decisions based on reports from financial analysts, participants ranked competence, trust, and likelihood of buying future reports higher for accurate versus inaccurate financial analysts (Kadous, Mercer, & Thayer, 2009). Also, compared to patients asked to imagine receiving an accurate test result, mammography patients asked to imagine receiving a false

positive breast cancer test result reported reduced trust and a greater likelihood of delaying mammography in the future (Kahn & Luce, 2003). Even preschoolers demonstrate reduced trust in an inaccurate informant compared to an accurate informant (Pasquini et al., 2007; Ronfard & Lane, 2018).

1.2 Perception of outcome

Inconsistency in information may also impact peoples' perception of a situation and the decisions that they base on it. Therefore, another important question about inconsistent information is whether people use both pieces of information to estimate what will happen. There is some research on this topic. When presented with inconsistent financial advice from multiple advisors (Budescu and Yu, 2007), or inconsistent forecasts from multiple forecasters (Su, Burgeno & Joslyn, 2021), participants' own estimates appear to be a simple average, suggesting that all the advice was considered and weighted equally. 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. In the context of sequential forecasts from the same source where more recent forecasts are more likely to be more reliable, it can be argued that the most recent forecast should be regarded as a replacement for the first.

Another way in which inconsistency may impact peoples' perceptions and subsequent decisions, is by signaling the extent of uncertainty in a situation. That is, when information is inconsistent, people may infer greater uncertainty. In other words, they may expect a greater range of outcomes when forecasts are uncertain.

1.3 Decision making

Furthermore, it is also important to consider how inconsistency may impact decisions. In the same simultaneous forecast study described above (Su, Burgeno & Joslyn, 2021), perhaps

because they perceived the situation to have greater uncertainty, participants made more cautious decisions (taking protective action more) when faced with inconsistent forecasts. This suggests that inconsistencies provided useful, decision-relevant information. However, it is unclear whether this is true of inconsistencies in sequential information from a single source.

1.4 Methods

Therefore, a primary goal of this project was to test the relative effects of forecast inconsistency and forecast inaccuracy on trust with sequential forecasts from the same source. I also investigated the effects of forecast inconsistency on participants' own outcome estimates, their uncertainty expectations (the range of outcome estimates participants would not be surprised by), and decisions.

All four experiments in the present research used a school closure paradigm in which participants were instructed to take on the role of a school administrator tasked with advising schools, according to a sequence of snow accumulation forecasts, whether they should stay open or close due to an impending snowstorm. They were instructed to advise schools to stay open if they expected less than six inches of snow accumulation, and to advise closing if they expected six or more inches of snow. Trials consisted of four screens presented sequentially: a Monday forecast for Wednesday, a Tuesday forecast for Wednesday, a decision screen where participants would indicate their school closure decision, and a screen with the Wednesday observed snow accumulation where participants would indicate how much trust they had in the forecasts they had just seen (see Figure 1).

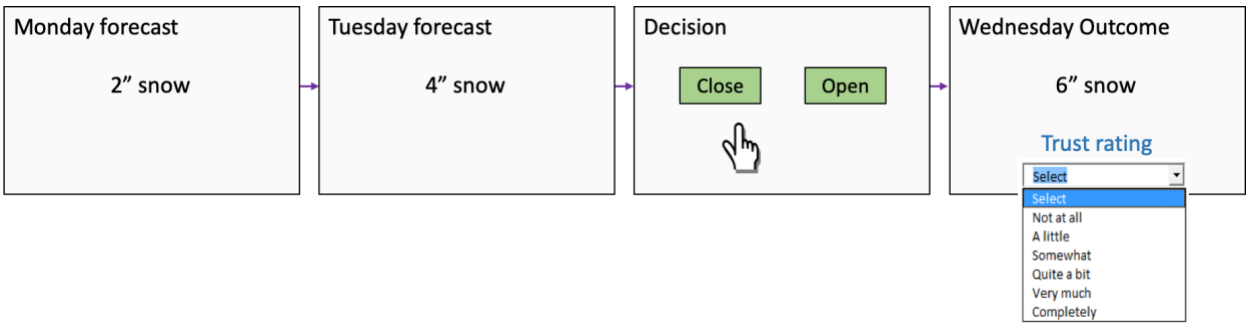

Figure 1. Trial events.

I utilized two main approaches to achieve the above goals. First, in order to ensure the effects observed were due to forecast inaccuracy or inconsistency, in Experiments 1-3 it was necessary to control a number of forecast characteristics (see Table 1). In particular, I attempted to maintain a realistic range of forecasted and observed values for the Pacific Northwest (4-7"), where the studies were conducted, because large levels of snow accumulation are unusual. If the snow accumulation forecast is unrealistic (e.g., 15 inches) then that could cause distrust for that reason, acting as an extraneous variable or in the worst case, a confound. Another potential extraneous variable was the magnitude of inconsistencies and inaccuracies. Therefore I also attempted to maintain equal magnitudes of inaccuracies and inconsistencies. Similarly, inconsistencies and inaccuracies may only be of considerable consequence to users if the difference crosses the threshold for decision making (6 inches). That is, because that would change the (value of the) decision they would make. For example, consider receiving an initial forecast of 3 inches, and an updated forecast of 5 inches as opposed to an initial forecast of 4 inches and an updated forecast of 6 inches. When inconsistency does not cross the decision threshold, the two forecasts indicate the same decision (stay open in this case) but when inconsistency crosses the decision threshold it may be less clear what decision one should make. Therefore, to eliminate the possibility of threshold crossing as an alternate explanation for the effects of inconsistencies or inaccuracies, and to make sure that both variables had the maximum

impact, I also attempted to have all inaccuracies and inconsistencies cross the 6-inch decision threshold. Likewise, since inaccuracy was a binary variable defined by the relationship between Forecast 2 and the Wednesday observed accumulation, allowing Forecast 1 to match the Wednesday observed accumulation may have obscured or undermined the effect of inaccuracy in the sense that it would have made forecasts at least partially accurate. Therefore, I attempted to make both Forecast 1 and Forecast 2 inaccurate for all inaccurate trials. Again it was not possible to control for all of these forecast characteristics in a single experiment; therefore, uncontrolled characteristics were traded across three otherwise tightly controlled experiments. The answers to my research questions (1-3 described below) regarding whether inconsistency impacts trust, how its impact on trust compares to that of inaccuracy, and whether it impacts expectations and decision making, are contained in the combined results of Experiments 1-3 reported in the published paper presented in Chapter 2.

Table 1

Forecast Characteristics By Experiment

	E1	E2	E3	E4
Balanced Forecast Types	✓	✓	✓	✗
Range of Values: 4-7”	✓	✗	✓	✗
Monday Values ≠ Wednesday Values	✓	✓	✗	✗
2” Magnitudes	✗	✓	✓	✗
Cross 6” Decision Threshold	✗	✗	✓	✗

½ Ascending, ½ Descending	✓	✓	✓	✗

While this strategy improved internal validity for the effects of forecast inaccuracy and inconsistency, it was at a tradeoff with their external validity. That is, unlike the forecasts that we experience in real life, the forecast stimuli used in these experiments had a limited range of values and subsequently limited magnitudes of inaccuracy and inconsistency such that they were essentially binary variables (e.g., either 0 or 2 inches). These experiments showed that the decrement in trust due to inaccuracy was greater than that of inconsistency. These experiments also showed that inconsistencies increased uncertainty expectations (operationalized as the range between participants' minimum and maximum estimates) and closure decisions (making decisions more cautious). Next, I tested whether these effects extend to naturalistic forecasts with real historic forecast data in which the degree of inconsistencies and inaccuracies varied naturally with Experiment 4 (presented in Chapter 3). In addition to trust, I also investigated the effects of inconsistency on what participants thought the outcome would be, uncertainty expectations, and decision making. That is, all experiments addressed the following primary research questions:

Research Question 1: Does forecast inconsistency reduce trust?

Research Question 2: Does inaccuracy or inconsistency cause a greater reduction in trust?

Research Question 3: Does inconsistency impact outcome expectations, uncertainty expectations, or decision making?

1.5 Uncertainty estimates

Additionally, there may be strategies to preserve trust in the context of inconsistency. For instance, perhaps providing explicit forecast uncertainty estimates can attenuate reductions in trust due to inconsistency. Prior research has demonstrated that uncertainty estimates tend to attenuate the reduction in trust due to inaccuracy. For example, in the road salting study described above, uncertainty estimates significantly reduced the negative impact of forecast inaccuracy on trust and decision making (Joslyn & LeClerc, 2012). In particular, participants rated trust higher when provided the probability of witnessing temperatures at or below freezing (the decision threshold), in addition to single-value forecasts, compared to those who received only single-value low temperature forecasts (Joslyn & LeClerc, 2012). Another study found that including uncertainty estimates preserved trust more than reducing false alarm rates (LeClerc & Joslyn, 2015). However, no prior work had attempted to evaluate whether uncertainty estimates similarly preserve trust in the face of inconsistencies. Therefore, I tested this question in two studies, Experiments 1 and 4. Experiment 1 tested this question but did not find an effect with unreliable probabilistic forecasts. Here reliability refers to the alignment of the forecasted probability and the experienced frequency of a given event. That is, in Experiment 1, forecasted exceedance probabilities, which ranged from 30-60%, did not align with the experienced frequency of observations at or above the 6 inch decision threshold. Instead, for every given probability the frequency of observing 6 or more inches was held constant at 50%. It was not possible to adjust forecasted probabilities in Experiment 1 to be reliable due to the tight control of forecast characteristics. All studies prior to this dissertation work showing advantages for probabilistic forecasts (Joslyn & LeClerc, 2012; LeClerc & Joslyn, 2015; Grounds, LeClerc, & Joslyn, 2018; etc.) had tested reliable forecasts (e.g., a 40% chance of a threshold observation was predicted, and the experienced frequency of observations at or above the threshold was 2 out

of 5 times). I did not expect that participants would detect unreliability over so few trials but indeed, it appears that the unreliability may have eliminated potential benefits of uncertainty estimates for trust. Therefore, a primary contribution of Experiment 4 is that it retested this question with reliable probabilities. It is the first study designed specifically to test whether providing reliable, naturalistic uncertainty estimate forecasts can attenuate reductions in trust due to inconsistency. In this naturalistic forecast experiment I also explored how inconsistencies and uncertainty estimates might impact outcome expectations, uncertainty expectations, decision making, and decision quality. That is, Experiment 4 addresses the following questions:

Research Question 4: Can including an uncertainty estimate attenuate the negative effect of inconsistency on trust?

Research Question 5: How do naturalistic inconsistencies and uncertainty estimates impact trust, outcome expectations, uncertainty expectations, decision making, and decision quality?

This dissertation is divided into two major parts which are presented in Chapters 2 and 3. Chapter 2 consists of Experiments 1-3 which used tightly controlled forecast characteristics to test the relative effects of forecast inconsistency and inaccuracy on trust. Chapter 3 consists of Experiment 4 which utilized forecast stimuli based on real storm data to test whether the pattern of effects observed in tightly controlled experiments extends to naturalistic forecast stimuli, and to different magnitudes of inaccuracy and inconsistency. In addition, Experiment 4 extends the prior research by examining whether there are effective strategies that can be taken to preserve trust in contexts of inaccuracy and inconsistency. Namely, I test whether providing an explicit uncertainty estimate, in addition to a deterministic forecast, preserves user trust in the forecast.

Chapter 2. EXPERIMENTS 1-3

The Impact of Weather Forecast Inconsistency on User Trust

Jessica N. Burgeno and Susan L. Joslyn

University of Washington

Author Note

Jessica N. Burgeno and Susan L. Joslyn, Department of Psychology, University of Washington.
Correspondence concerning this article should be addressed to Jessica N. Burgeno, Department of Psychology, University of Washington, Seattle, WA 98103.
Email: jburgeno@uw.edu

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.

The Impact of Forecast Inconsistency on User Trust

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, Morss, & Demuth, 2009; Wilson & 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 & 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 & 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” to 6”) or downward (6” to 4”) revision will continue moving in the same direction into the future (Hohle & Teigen, 2015; Erlandsson, Hohle, Løhre, & Västfjäll, 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, Rantilla, Yu, & Karelitz, 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 & Yu, 2007). It is important to note however that these are simultaneous forecasts from different sources that appear to "disagree" with one another (Løhre, Sobkow, Hohle and Teigen, 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 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 & 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, Scolobig, & Patt, 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 a sacrifice to accuracy. For example, in October 2016 historic and destructive

winds were forecasted for Saturday, October 15th 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 & 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, Mercer & Thayer, 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 & Luce, 2003). Even preschoolers have been found to trust accurate more than inaccurate informants (Pasquini, Corriveau, Koenig, & Harris, 2007), and adjust their trust according to subsequent accuracy (Ronfard & Lane, 2018).

Therefore, because consistency is widely advocated and the maintenance of consistency could be at the sacrifice to 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 lab-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 twelve weeks, for a total of twenty-four 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 6 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 whom, 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, pre-testing 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 Figure 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 inches. These values were 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-inch difference between the second (Tuesday) forecast and the observed value. All inaccurate trials crossed the 6-inch 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-inch 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-inch 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-inch 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 inches 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-inch decision threshold. Therefore, in Experiment 1 the inconsistencies in inaccurate trials were 1 inch while in accurate trials they were 2 inches (we return to this issue in the discussion). In order to control for any effects that might result from deducing trends over the two forecasts (Hohle & Teigen, 2015, 2018; Maglio & 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-inch discrepancy between the second forecast and observed value and did not cross the 6-inch closure threshold. Filler trial values were lower (2 or 3 inches) or higher (8 or 9 inches) than values for experimental trials (4 to 7 inches).

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 30-60%, 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-inch 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-groups 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.

In order 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, SD=0.82) significantly higher than their trust in inconsistent forecasts (M=3.07, SD=0.88), independent of accuracy, $F(1, 361)=17.35$, $p<.001$, Cohen's $d=.24$ (see Figure 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<.001, Cohen's d=.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=.003, Cohen's d=.13*. Post-hoc, Bonferroni corrected paired comparisons confirmed that consistency did not have a significant effect on trust when forecasts were inaccurate, *t(362)=1.69, p=.092*, although the effect was significant when they were accurate, *t(362)=4.94, p<.001*.

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 (at a potential sacrifice to 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<.001, Cohen's d=.30*. This analysis is especially relevant considering that these trial types featured inaccuracies and inconsistencies with equal magnitudes (2 inches).

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 pre-outcome 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< .001,* the effect size was small, $Cohen's\ d$=.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 .83 unit increase in snow accumulation estimates, $\beta=.47$, $p<.001$ while a one unit increase in the first forecast value predicted only a .11 unit increase, $\beta=.06$, $p<.001$. 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, $F(2, 2891)=421.89$, $p<.001$, $R^2=.23$[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), $t(362)=3.97$, $p<.001$. 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, $r = -.98$, $p=.006$), 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 Experiment 1

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-inch 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-inch, 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-inch threshold. In addition, and somewhat surprisingly, the effect of inconsistency on pre-outcome 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 inches. In addition, Experiment 2 tested 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 inches) allowing for 2-inch 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-groups 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 post outcome trust indicated that (independent of accuracy) trust ratings for consistent forecasts (M = 3.32, SD = .07) were significantly higher than for inconsistent forecasts (M = 3.12, SD = .07), *F(1, 160)=12.20, p=.001, Cohen's d=.34* (see Figure 3). In addition, (independent of consistency) trust ratings for accurate forecasts (M =3.42, SD = .06) were significantly higher than for inaccurate forecasts (M =3.03, SD = .07), *F(1,160)=54.58, p<.001, Cohen's d=.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=.14, Cohen's d=.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=.020, *Cohen's d=.19*.

In addition, pre-outcome 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= .001, Cohen's d=0.24.* As with Experiment 1, the effect was smaller than the post outcome effect of inconsistency (*Cohen's d*=.34), and smaller than the effect of inaccuracy on post outcome trust (*Cohen's d*=.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, $\beta=.78$, $p<.001$, while a one unit increase in the first forecast value predicted only a 0.04 unit increase in snow accumulation estimates, $\beta=.06$, $p<.001$. Overall, the two predictor model accounted for 67% of the variance in snow accumulation estimates, $F(2, 2589)=2594.99$, $p<.001$, $R^2=.67$.

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]). In order 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 inches (below threshold) and 7 inches (above threshold). This was done because decisions for values at or near the 6-inch 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 & 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= .61, SD=.48) than on consistent forecasts (M= .59, SD=.49), $F(1, 160)= 9.83$, $p=.002$, *Cohen's d=.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<.001, Cohen's d=.21* (see Figure 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=.93, SD=.25) than below (M=.16, SD=.37) the threshold, *F(1, 160) = 1528.52, p<.001, Cohen's d=3.74.*

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 Experiment 2

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 pre-outcome 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.

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 inches.

However, one confound remained. Although all inaccuracies and inconsistencies were equal in magnitude in Experiment 2, only half of inconsistencies crossed the 6-inch 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 inches 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 inches) *and* crossed the 6-inch 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-inch inconsistencies could cross the 6-inch decision threshold (e.g., first forecast: 4 inches, second forecast: 6 inches, observed: 4 inches). 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.

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 post-outcome 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, $F(1, 156) = 42.66$, $p<.001$, *Cohen's d=.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, $F(1, 156)= 66.68$, $p<.001$, *Cohen's d=.79* (see Figure 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<.001, Cohen's d=.22*. Bonferroni corrected paired comparisons revealed that although the difference was larger for accurate than inaccurate forecasts, it was significant for both (inaccurate, *t(157)=3.27, p=.001*; accurate trials, *t(157)=7.83, p<.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=.244, Cohen's d= .10*. Again, the effect on pre-outcome trust was small. A paired samples t-test revealed significantly higher pre-outcome trust ratings for consistent (M= 3.33, SD=0.77) than inconsistent trials (M= 3.01, SD=0.81), *t (157) = 6.08, p< .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 .79 unit increase in estimated snow accumulation,

β= .81, p<.001

, while a one unit increase in the first forecast value predicted a .13 unit increase in estimated snow accumulation,

β= .13, p<.001

. Overall, the two predictor model explained a significant proportion of the variance in snow accumulation estimates[4], $F(2, 2524) = 2623.08$, $p<.001$, $R^2 = .68$.

c. Discussion Experiment 3

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-inch 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 pre-outcome 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 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.

In order 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 categorial inconsistencies (e.g., inconsistent forecast: a "dusting of snow" to "several inches", Losee & 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 & 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 post outcome 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 pre-outcome trust rating. Here too, in all three experiments, consistency had only a small effect, smaller than the effect on post outcome trust and much smaller than the effect on trust due to inaccuracy. This contradicts the intuition that the diagnostic relevance of consistency (Falk & Zimmermann, 2017) should be *greater* in the absence of accuracy information. One possible explanation for the smaller impact of inconsistency pre- than post-outcome, 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's 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 & 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, Smallman & Moore, 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 extra-

experimental experience with real weather forecasts about which people have many, often valid intuitions (Morss, Demuth, Lazo, 2008; Joslyn & Savelli, 2010; Savelli & 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, unlike simultaneous predictions from separate sources that tend to be weighted equally (Budescu & 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 a sacrifice to consistency, can preserve trust in the information source as well as provide users with higher quality decision-relevant information.

Acknowledgements

This research was supported by the National Science Foundation under grant 1559126.

Data for all experiments can be found at

https://osf.io/rzbcw/?view_only=4cea5204371e48beb6ba2e65d8bba34d.

Appendix A

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. We also provide an analysis 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.

Indeed, snow accumulation estimates for ascending forecasts were significantly larger than for descending forecasts in both Experiment 1, $t(362)= 13.86$, $p<.001$, and 3, $t(157)= 21.57$, $p<.001$ (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, $t(362)= 24.43$, $p<.001$, and 3, $t(157)= 13.46$, $p<.001$ (see 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.

Furthermore, the difference in trend validated and trend contradicted trials failed to reach significance, $t(362)= 1.23$, $p=.22$ (see Table A3), suggesting that our primary findings aren't explained by the trend effect.

References

Budescu, D. V., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes, 90*(1), 178-194.

Budescu, D. V., & Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153-177.

Drabek, T. E. (1999) Understanding disaster warning responses. *The Social Science Journal*, *36*, 515–523.

Elder, K., Xirasagar, S., Miller, N., Bowen, S. A., Glover, S., & Piper, C. (2007). African Americans' decisions not to evacuate New Orleans before Hurricane Katrina: A qualitative study. *American Journal of Public Health*, *97*(Supplement 1), S124-S129.

Falk, A., & Zimmermann, F. (2017). Consistency as a signal of skills. *Management Science*, 63(7), 2197-2210.

Erlandsson, A., Hohle, S. M., Løhre, E., & Västfjäll, D. (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. *Journal of Applied Social Psychology*, *48*(11), 618-633.

Hohle, S. M. & Teigen, K. H. (2015). Forecasting forecasts: The trend effect. *Judgment and Decision Making*, *10*(5), 416-428.

Hohle, S., & Teigen, K. (2018). When probabilities change: Perceptions and implications of trends in uncertain climate forecasts. *Journal of Risk Research*, 1-15.

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126.

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, *17*(2), 180-195.

Kadous, K., Mercer, M., & Thayer, J. (2009). Is there safety in numbers? The effects of forecast accuracy and forecast boldness on financial analysts' credibility with investors. *Contemporary Accounting Research*, *26*(3), 933-968.

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: mammography adherence following false-alarm test results. *Marketing Science*, *22*(3), 393-410.

Lazo, J. K., Morss, R. E., & Demuth, J. L. (2009). 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bulletin of the American Meteorological Society*, *90*(6), 785-798.

Løhre, E., Sobkow, A., Hohle, S. M., & Teigen, K. H. (2019). Framing experts'(dis) agreements about uncertain environmental events. *Journal of Behavioral Decision Making*, *32*(5), 564-578.

Losee, J. E., & Joslyn, S. (2018). The need to trust: how features of the forecasted weather influence forecast trust. *International journal of disaster risk reduction*, *30*, 95-104.

Maglio, S. J., & Polman, E. (2016). Revising probability estimates: Why increasing likelihood means increasing impact. *Journal of Personality and Social Psychology*, *111*(2), 141-158.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the US public. *Weather and Forecasting*, *23*(5), 974-991.

National Oceanographic and Atmospheric Administration. (2016). *Risk communication and behavior: Best practices and research findings*. Silver Spring, MD : U.S.

Oonk, H. M., Smallman, H. S., & Moore, R. A. (2001). Evaluating the usage, utility and usability of web-technologies to facilitate knowledge sharing. In Proceedings of the Command and Control Research & Technology Symposium.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology, 43*(5), 1216.

Perry, R.W. and Green, M.R. (1982) The role of ethnicity in the emergency decision-making process. *Sociological Inquiry*, *52*, 306–334.

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

Ronfard, S., & Lane, J. D. (2018). Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. *Child Development*, *89*(2), 414-429.

Savelli, S., & Joslyn, S. (2012). Boater safety: Communicating weather forecast information to high-stakes end users. *Weather, Climate, and Society, 4*(1), 7-19.

Smithson, M. (1999). Conflict aversion: Preference for ambiguity vs conflict in sources and evidence. *Organizational Behavior and Human Decision Processes*, *79*(3), 179-198.

Weyrich, P., Scolobig, A., & Patt, A. (2019). Dealing with inconsistent weather warnings: Effects on warning quality and intended actions. *Meteorological Applications*, 1-15.

Wilson, L. J., & Giles, A. (2013). A new index for the verification of accuracy and timeliness of weather warnings. *Meteorological Applications*, *20*(2), 206-216.

Table 1

Inches of Snow Forecasted and Observed by Experiment and Experimental Trial Type

	Accurate			Inaccurate				
	First Forecast	Second Forecast	Observed Outcome	First Forecast			Second Forecast	Observed Outcome
					E1	E2 E3		
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
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

Note. Bold values highlight differences in forecast values across experiments. All other values were the same across all experiments. Filler trials not included. Participants were advised to close schools if they expected 6 or more inches of snow.

Table 2

Mean Trust Ratings by Experiment and Trial Type

Experiment		Accurate	Inaccurate	Consistent	Inconsistent	Accurate Consistent	Accurate Inconsistent	Inaccurate Consistent	Inaccurate Inconsistent
1	**Mean Pre**			**3.19**	**3.05**				
	Std dev			**.79**	**.82**				
	Mean Post	**3.35**	**2.93**	**3.21**	**3.07**	**3.45**	**3.25**	**2.96**	**2.90**
	Std dev	.79	.904	.82	.88	.86	.91	1.01	1.01
2	**Mean Pre**			**3.28**	**3.09**				
	Std dev			**.81**	**.74**				
	Mean Post	**3.42**	**3.03**	**3.32**	**3.12**	**3.54**	**3.29**	**3.10**	**2.96**
	Std dev	.78	.93	.85	.88	.86	.90	1.05	1.01
3	**Mean Pre**			**3.33**	**3.01**				
	Std dev			**.77**	**.81**				
	Mean Post	**3.45**	**3.00**	**3.41**	**3.05**	**3.70**	**3.21**	**3.11**	**2.89**
	Std dev	.80	.90	.87	.83	.91	.88	1.08	.91

Table 3

Post Decision Trust Analyses: Test Statistics and Effect Sizes by Experiment and Effect

Experiment		Accuracy	Consistency	Accuracy x Consistency
1	F	124.83***	17.35***	8.78**
	Cohen's d	.79	.24	.13
2	F	52.18***	12.59**	1.94
	Cohen's d	.69	.34	.09
3	F	66.65***	42.41***	13.72***
	Cohen's d	.79	.61	.22

*p<.05; **p<.01; ***p<.001

Table A1

Mean and SD Snow Accumulation Estimates by Consistent and Inconsistent Conditions

Snow Estimate	Experiment 1 (N=363)		Experiment 3 (N=158)	
Consistent	M=5.42, SD=1.21 Trials=8		M=5.45, SD=1.22 Trials=8	
Inconsistent	Ascending M=6.05 SD=1.02 Trials=4	Descending M=4.96 SD=2.49 Trials=4	Ascending M=6.14 SD=0.78 Trials=4	Descending M=4.78 SD=0.87 Trials=4

Note. Inconsistent conditions are broken down by ascending and descending categories.

Table A2

Mean and SD School Closure Decisions by Consistent and Inconsistent Conditions

% Closed	Experiment 1(N=363)		Experiment 3(N=158)	
Consistent	M=0.57, SD=0.50 Trials=8		M=0.57, SD=0.50; Trials=8	
Inconsistent	Ascending M=0.78 SD=0.42 Trials=4	Descending M=0.40 SD=0.49 Trials=4	Ascending M=0.78 SD=0.50 Trials=4	Descending M=0.42 SD=0.41 Trials=4

Note. Inconsistent conditions are broken down by ascending and descending categories.

Table A3

Experiment 1 (N=363) Mean and SD Trust Ratings by Accuracy and Consistency

Trust	Accurate	Inaccurate	
Consistent	M=3.45, SD=1.22; Min=0, Max=6; Trials=4	M=2.96, SD=1.34; Min=0, Max=6; Trials=4	
Inconsistent	M=3.25, SD=1.19; Min=0, Max=6; Trials=4	Verified M=2.93, SD=1.29; Min=0, Max=6; Trials=2	Contradicted M=2.86, SD=1.35; Min=0, Max=6; Trials=2

Note. Inconsistent/Inaccurate conditions are broken down by trend verified and trend contradicted categories.

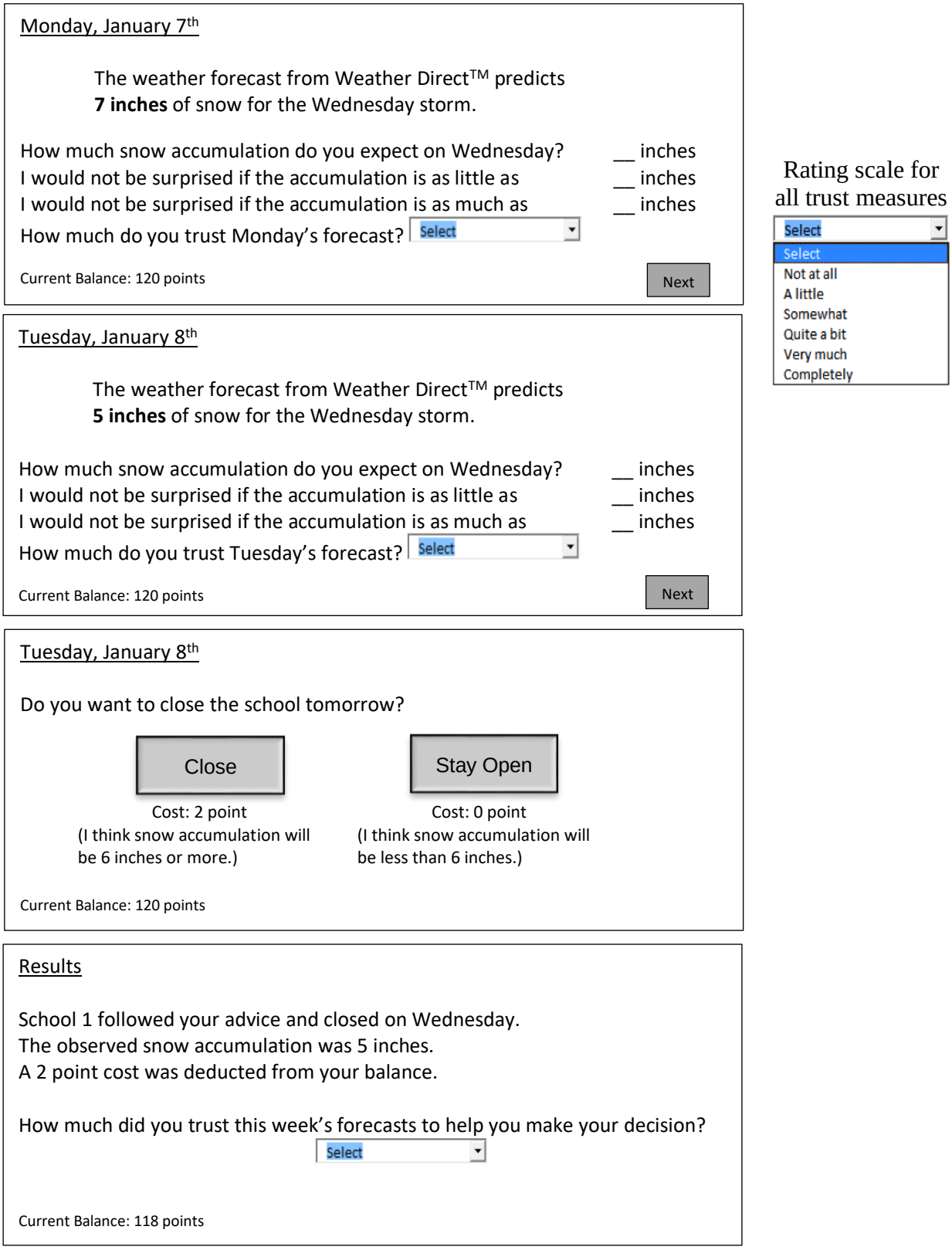

Figure 1. Screens Shown in a Single Trial in Order from Top to Bottom

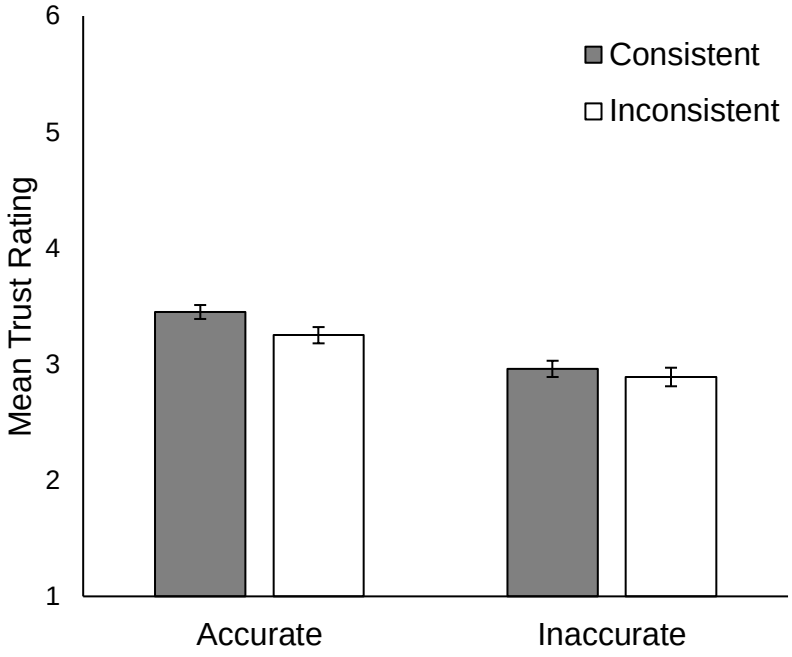

Figure 2. Experiment 1 Trust Ratings by Accuracy and Consistency.
Note. Error bars show 95% CI.

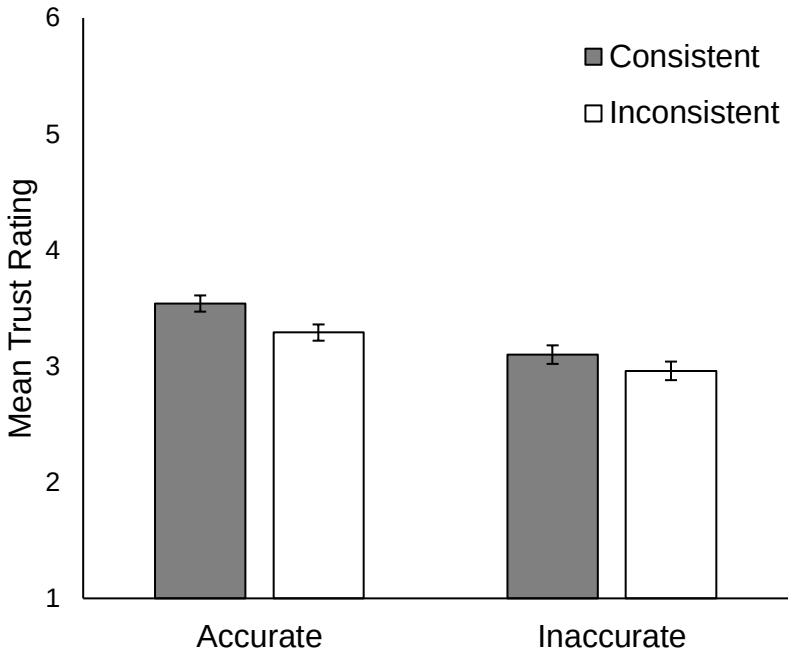

Figure 3. Experiment 2 Trust Ratings by Accuracy and Consistency.
Note. Error bars show 95% CI.

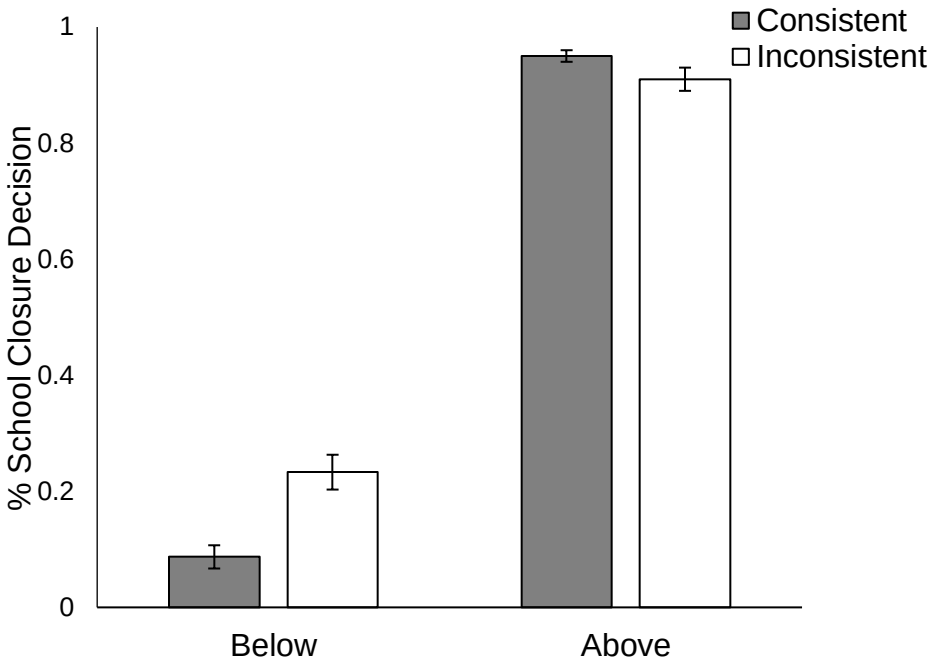

Figure 4. Experiment 2 Percent Closed by Threshold Orientation and Consistency.
Note. Error bars show 95% CI.

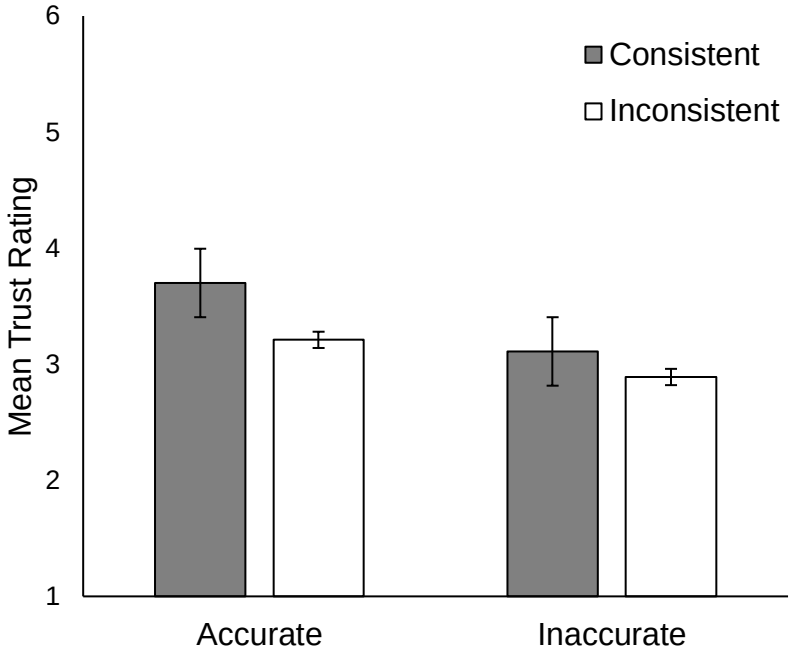

Figure 5. Experiment 3 Trust Ratings by Accuracy and Consistency.
Note. Error bars show 95% CI.

Chapter 3. EXPERIMENT 4

The Impact of Forecast Inconsistency and Probabilistic Forecasts on Users' Trust and Decision-Making

Jessica N. Burgeno and Susan L. Joslyn

University of Washington

Author Note

Jessica N. Burgeno and Susan L. Joslyn, Department of Psychology, University of Washington. Correspondence concerning this article should be addressed to Susan L. Joslyn, Department of Psychology, University of Washington, Seattle, WA 98103.
Email: susanj@uw.edu

Abstract

When forecasts for a major weather event begin days in advance, updates may be more accurate but can be inconsistent with the original forecast. Resulting inconsistency may reduce user trust. However some evidence suggests that the reduction in trust due to inaccuracy is greater (Burgeno & Joslyn, 2020). The experiment reported here tested the impact of both inconsistency and inaccuracy on trust and decision quality. It also investigated whether providing probabilistic forecasts along with single value forecasts attenuates the hypothesized negative effects of either forecast inconsistency or inaccuracy. Participants rated their trust in snow accumulation forecasts that were based on historical records and used them to make school closure decisions. Half of participants received single-value forecasts and half also received the probability of observing 6 or more inches of snow (the decision threshold in the assigned task). In line with previous, forecast inaccuracy was detrimental to trust although probabilistic forecasts attenuated the effect. Further, the inclusion of probabilistic forecasts allowed participants to make better decisions from an economic perspective. Surprisingly, in this study, inconsistency increased, rather than decreased, trust in forecasts. Perhaps because inconsistency also provided participants with useful information, alerting them to forecast uncertainty and leading them to make more cautious decisions. This work has important implications for many practical settings, suggesting that both probabilistic forecasts and forecast inconsistency provide useful information to decision makers. Therefore, if forecasters have well calibrated uncertainty estimates or newer, more reliable information, they are advised to share such information.

The Impact of Forecast Inconsistency and Probabilistic Forecasts on Users' Trust and Decision-Making

1. Introduction

Forecasts for major weather events often begin days in advance. The weather models upon which forecasts are based update frequently and generally grow more accurate on average as lead times decrease (Lazo et al., 2009; Wilson & Giles, 2013). However, meteorologists are often reluctant to update the forecasts provided to members of the public out of fear that inconsistency in subsequent forecasts (i.e., most recent forecast differs from prior forecast) will be confusing and negatively affect user trust. In fact, maintaining consistency in forecasts is considered best practice by some institutions, like the National Oceanic and Atmospheric Administration (NOAA, 2016). Yet, the choice to maintain consistency can be at a loss to accuracy (how closely the forecast matches the outcome).

Indeed, there is strong evidence that forecast inaccuracy reduces trust. For instance, in a study in which participants used overnight low temperature forecasts to make road salting decisions, they rated trust significantly higher and took protective action more often with low- compared to high-error forecasts (Joslyn & LeClerc, 2012). Similarly, in a study in which participants used reports from financial analysts to make investment decisions, participants rated competence, trust, and likelihood of buying future reports higher for accurate compared to inaccurate financial analysts (Kadous, Mercer, & Thayer, 2009). Similarly, when asked to imagine receiving false positive breast cancer test results, mammography patients reported reduced trust and being more likely to delay future mammography compared to those who imagined receiving an accurate initial test result (Kahn & Luce, 2003). Even preschoolers show

reduced trust in inaccurate relative to accurate informants (Pasquini et al., 2007; Ronfard & Lane, 2018).

There is also recent evidence that speaks to the effect of inconsistency in predictions on trust and decision making. For example, one study that manipulated forecast consistency in sequential thunderstorm and snow forecasts from a single source found that consistent (relative to inconsistent) forecasts led to greater trust (Losee & Joslyn, 2018). There is also research comparing the impact of inconsistency to that of inaccuracy, suggesting that sequential forecast inconsistency negatively affects user trust, but that inaccuracy has a larger negative effect on trust (Burgeno & Joslyn, 2020). In these experiments, participants based school closure decisions on snow accumulation forecasts (e.g., Monday forecast: 4 inches of snow on Wednesday) from a single source, 2-days and 1-day in advance of an anticipated storm. Not only was inaccuracy more detrimental to trust in the forecast but inconsistency appeared to provide useful information. It increased uncertainty expectations and led to more conservative closure decisions. An inaccuracy by inconsistency interaction effect suggested that differences in trust due to inconsistency shrank when forecasts were inaccurate. In other words, the reduction in trust due to inaccuracy was substantial to the extent that inconsistency had little additional impact.

At least part of the reason that inconsistency is less detrimental to trust in sequential forecasts may be the fact that, when forecasts are inconsistent, people understand that the most recent forecasts is more likely to be more accurate and regard it as a replacement for an earlier forecast. Indeed, prior research on sequential forecasts suggest that participants' best estimates were more heavily influenced by recent forecasts (1-day in advance) than initial forecasts (2-

days in advance), suggesting that participants expected the most recent forecast to be more reliable (Burgeno & Joslyn, 2020).

The relative effects of inconsistency and inaccuracy have also been compared in an experiment based on snow forecasts from 2 different sources both provided at the same time, one day in advance of an anticipated storm. It revealed that while inaccuracy significantly reduced trust, inconsistency between the two sources did not (Su, Burgeno, & Joslyn, 2021). In fact, participants incorporated information from both sources equally in their outcome estimates and appeared to glean useful information from inconsistencies. That is, as with inconsistent sequential forecasts (Burgeno & Joslyn, 2020), inconsistencies led participants to infer greater uncertainty and to make more cautious decisions. Therefore, inconsistency appears to be less problematic for trust than inaccuracy in both sequential forecasts coming from the same source and simultaneous forecasts from different sources, and it may provide useful information.

Although this research is informative, it's important to note that to identify and compare these effects, all of the experiments cited above used highly controlled forecast stimuli, limiting the range of forecasts values and closely matching the degrees of inaccuracy, inconsistency (about 2 inches) and proportions (50%) of each. That begs the question, will the same effects be observed in realistic forecast situations in which forecasts, outcomes, inconsistencies and inaccuracies take on a wider range of values and vary naturally? The experiment reported here was designed to evaluate that question.

The other question this work was designed to answer is whether there is a benefit to adding an uncertainty estimate to inconsistent forecasts. Although forecast inconsistency may reduce trust in some situations, it may be possible to preserve trust in the face of inconsistency by adding explicit uncertainty estimates to the forecast as has been shown with inaccuracy. For

example, in the road salting study (Joslyn & LeClerc, 2012) mentioned above, uncertainty estimates reduced the negative effects of forecast inaccuracy on both trust and decision making. When provided the probability of observing temperatures at or below the decision threshold (in addition to single-value forecasts) participants rated trust higher than those who received single-value low temperature forecasts alone. They also made better decisions from an economic perspective (Joslyn & LeClerc, 2012). In another study, providing participants with probabilistic forecasts preserved trust to a greater degree than lowering false alarm rates. Including probabilistic forecasts also increased compliance with weather warnings (LeClerc & Joslyn, 2015). In yet another set of studies, probabilistic forecasts added to flood warnings enhanced subjective understanding of flood likelihood and reduced recency biases compared to a return period expression (e.g., 10 year flood) and to a no information control group (Grounds, LeClerc, & Joslyn, 2018). Thus, a growing body of evidence suggests that laypeople can use explicit probabilistic information and that it may offer a number of benefits in the decision-making process. The experiment reported here was designed to test whether including explicit uncertainty estimates in forecasts preserves trust in the face of forecast inconsistency as well.

In sum the experiment reported here was designed to test whether the reduction in trust due to forecast inconsistency extends to inconsistency values that vary naturally and if so, whether the reduction in trust is attenuated by including uncertainty estimates in the forecast. It also tested the impact of these factors on participants own outcome estimates and decision quality. This experiment makes use of the school closure paradigm discussed above. Participants' goal was to decide, based on a sequence of snow forecasts taken from historical records, when it was appropriate to close schools due to a snowstorm using a 6-inch or more

accumulation rule. Half of participants received probabilistic forecasts in addition to the single-value snow accumulation amount to determine the impact on trust and decision quality.

We hypothesized that inconsistency and inaccuracy would reduce trust and that probabilistic forecasts would enhance trust. We also hypothesized that more recent forecast, in inconsistent pairs, would have a greater impact on participants' accumulation estimates. We predicted that inconsistency would be interpreted as indicating greater uncertainty. Furthermore, we hypothesized that probabilistic forecasts would enhance decision quality and attenuate the negative effects of forecast inconsistency and inaccuracy on trust. We also tested whether the positive effect of inconsistency on uncertainty expectations (the range of accumulation estimates participants would not be surprised by) would be affected by probabilistic forecasts. Hypotheses were preregistered on Open Science Framework and can be viewed at https://osf.io/dv6j8.

a. Method

1) Participants. A total of 419 University of Washington psychology students participated for course credit and the opportunity to earn a cash bonus. After executing data cleaning procedures (described below), data from 398 participants (62% female, mean age=19.5) remained and were included in the analyses below.

2) Apparatus. The experiment was programed in Excel Visual Basic and administered on standard desktop computers.

3) Procedure. Participants were tested in groups of about eight. They first gave informed consent and provided their age and gender. Next participants read, and listened to, instructions that explained the task. Participants were asked to provide decision advice to schools regarding whether they should stay open or close due to an upcoming snowstorm.

Although several factors are considered when actual closure decisions are made, 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 for 65 schools across the region for each of two hypothetical winter periods, for a total of 130 trials. Each week was described as involving a different school district to encourage participants to regard the trials as independent of one another.

To encourage engagement with the task, participants began with a virtual budget of 332 points. Their goal was to retain as many points as possible. 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 was observed, a 6-point penalty was deducted from the score to reflect the risk of travel in dangerous road conditions. They earned a cash bonus for the ending point balance at the rate of $1 for every 32 points over 72 (final balance) points. A 332 initial point endowment was selected to discourage the simplistic and unrealistic strategy of recommending closure for every trial, which would result in a final balance at the payment threshold (72 points)[1].

For each trial, participants were to base their school closure decision on two snow forecasts for Wednesday, one issued on Monday (two days—48 hours—prior to the event) and one on Tuesday (one day—24 hours— prior to the event). The two snow accumulation forecasts for Wednesday were presented sequentially, centered on separate screens. The current weekday

[1] In particular, the initial endowment was calculated by multiplying the number of trials (130) by the cost of closing (2 points), and adding that product to the payment threshold, (130*2)+72=332. Allowing for a 72 point remaining balance was intended to maintain task engagement. That is, they still had points to play with as an incentive to remain engaged.

appeared in the upper left hand corner of each screen in bold font. In order to determine how the forecasts influenced participants own estimates, below each forecast, they were asked to provide the number of inches of snow they expected for Wednesday as well as the least (minimum estimate, "as little as") and greatest (maximum estimate, "as much as") number of inches that they would not be surprised by. Then, participants rated their trust in the forecast on a 6-point drop-down menu, from "Not at all" to "Completely". The 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 in the lower right-hand corner of the screen to advance to the next screen. Once the "next" button was pressed, participants could not go back and change answers on the previous screen. After the second forecast, participants were shown a decision screen. The current day (Tuesday) was shown in bold font in the upper left-hand corner of the screen. In order to provide their recommendation to the current school, participants pressed one of two buttons in the middle of the screen labeled "close" and "stay open." Text below each button reminded participants that close meant "I think snow accumulation will be 6 inches or more" and that stay open meant "I think snow accumulation will be less than 6 inches."

After submitting their school closure decision, a fourth screen appeared stating that the school followed their advice and either stayed open or closed. The observed snow accumulation on Wednesday 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 forecast using the same pull-down menu. In sum, each trial consisted of 4 screens: 1) Monday forecast for Wednesday, 2) Tuesday forecast for Wednesday, 3) Tuesday night school closure decision, and 4) Wednesday outcome. Then, the next trial began

with new set of forecasts and outcome, that pertained to a school in a different district. Participants performed four practice trials before the test trials began.

4) Stimuli.

i. Historical Forecast Data Set. The data upon which the snow accumulation forecasts (48 and 24 hours in advance), associated probabilities of 6 or more inches accumulation and the observed 24 hour snow accumulation outcomes were based, were obtained from the Eastern Region Headquarters of the National Oceanographic and Atmospheric Administration (NOAA)[2]. The original set of 160 complete forecasts pertained to a snowstorm that occurred in several locations over the eastern United States on February 9, 2017[3]. In the experiment we used 130 of these, treating each pair of forecasts and outcome as a separate event. All single value forecasts and observed accumulation amounts were rounded to the nearest inch. We maintained the approximate proportion (64 trials, 49%) of consistent forecasts (same single-value [4] for forecast 1 and 2) as in the original dataset. We increased the proportion of descending forecasts among those that were inconsistent. Because the original cases pertained to a single weather event in which the expected accumulation increased over time, there were very few descending forecast pairs (7 trials, 5%). This was problematic because some anecdotal evidence[5] suggests that descending forecasts are more likely to be altered to maintain consistency by forecasters. Therefore, cases were selected to increase the proportion of descending forecasts in the final forecast stimuli set (see forecast characteristics section below).

ii. Forecast Format. Half of participants received a single value forecast, while the other

[2] Of the 160 complete cases, 130 were used in the experiment
[3] Special thanks go to David B. Radell at NOAA and the National Weather Service for providing us with the forecast data.
[4] Of the exactly consistent trials, 20 out of 38 had the same probabilities
[5] Unpublished interviews with operational forecasters at National Weather Service Western Region, Seattle, WA.

half received the same single value and the 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"). We refer to the former as deterministic forecasts in that they imply an exact outcome (e.g., "…4 inches of snow"). Thus, other than the additional probability of observing 6 or more inches of snow, the forecasts and outcomes seen by both groups of participants were identical and presented in one of four fixed orders. Forecasted snow accumulation totals ranged from 0 to 17 inches (first forecast Mean=5.18, second forecast Mean=6.12), and observed values (M=5.21) ranged from 0 to 20 inches[6]. The probabilities of 6 or more inches ranged from 0-100% (first forecast Mean=35.45%, second forecast Mean=42.02%)[7].

iii. Calibration Procedures. Because it is important to first test the impact of well-calibrated probabilistic forecasts[8], especially for the most recent forecast used as the standard for accuracy (see section iv. below), some second forecasts were altered slightly so that forecasted probabilities for 6 or more inches of snow accumulation roughly matched the frequency of observing 6 or more inches of snow. A binning technique was used because there were very few cases at the exact same probability, precluding more conventional measures such as the Brier score (Brier, 1950). A bin was considered calibrated if the proportion of observed events with 6+ inches of accumulation fell within the probability range for that bin. For instance, Bin 2 ranged between 5-14% and contained 1 out of 10 trials (10%) in which 6 or more inches of snow accumulation was observed. See Appendix A.

---

[6] This was similar to the original data set, in which forecasted snow accumulation values ranged from 0 to 18 inches (first forecast Mean= 3.79, second forecast Mean=5.83), and observed values ranged from 0 to 20 inches (Mean= 4.46).

[7] In the original data set (which slightly under-forecasted the probabilities), the first forecast mean was 26.06% and the second forecast mean was 37.78%.

[8] Otherwise, null effects could be due to either the genuine lack of an affect or simply the lack of an affect for uncalibrated probabilities.

In the historical forecast data set, the proportion of outcomes at or above the 6 inch threshold was within a few percentage points of the bin boundaries in most cases. However, in the higher probability bins (65-74%,75-84%, and 85-94%), the proportions were as many as 25 percentage points higher than the upper bound of the bin suggesting a low bias in the forecasted probabilities for that day. Therefore, slight changes were made (cases were removed, duplicated, and/or modified) to perfect probabilistic forecast reliability while maintaining the basic characteristics of the historical forecast set.

iv. Forecast Characteristics. Thus, the forecast stimuli for this experiment consisted of 130 forecast pairs and outcomes that were based on historic data and retained the relevant characteristics of that data set. As a result, forecasts varied naturally in terms of both accuracy and consistency.

Accuracy was gauged relative to the second (Tuesday) forecast. By this standard, the proportion of exactly accurate forecasts was 22%, as with the historical forecast set. All inaccurate trials were inaccurate by 1 inch or greater. Inaccuracies ranged from -6 to 10 inches and had a mean inaccuracy of -.91 inches (similar to the original data set). Thus overall, inaccurate forecasts were biased high by just under an inch. Similar to the historical forecast set, twenty-five inaccurate trials (24% of inaccurate trials, 19% of all trials) had inaccuracies that crossed the 6 inch decision threshold (e.g., a second forecast of 7 inches and an observed snow accumulation of 5 inches).

Consistency was defined as an exact match between the Monday and Tuesday forecasts. There were 64 (49%) consistent trials, and 66 (51%) inconsistent trials in the final forecast stimuli. All inconsistent trials were inconsistent by 1 or more inches with a range of -8 to 8 inches and a mean of 2.86 inches. Of the inconsistent trials, 40 (61%) were ascending (values

increased from first to second forecast) and the remaining 26 (39%) trials were descending (values decreased from first to second forecast). Twenty-one inconsistent trials (32% of inconsistent trials, 16% of all trials) had inconsistencies which crossed the 6 inch decision threshold (e.g., a first forecast of 5 inches and a second forecast of 9 inches). This was a slightly smaller proportion of threshold crossing inconsistencies than in the highly controlled experiments where at least half of inconsistencies crossed the threshold (25% of all trials in each respective experiment). The average forecasted accumulation for inconsistent forecasts was slightly less (M=5.67 inches), than for consistent forecasts (M=6.19 inches) .

5) Design. A single factor (forecast format) between-participants design with 2 levels, deterministic, or probabilistic, was used. Participants were randomly assigned to one of these two conditions. Forecast values, the magnitude of inconsistency and inaccuracy, and the economically optimal decision (see closure decision analysis below) were also included as predictor variables. The outcome variables were trust rating, snow accumulation estimates, and participants' decisions about whether to close schools or not (closure decisions).

b. Results

Prior to conducting the main analyses, we eliminated participants who did not understand the task, were not paying attention or taking the task seriously. To this end, participant data were excluded if a) they provided a lower estimate for maximum than minimum estimate, or if b) their average best estimate, or c) highest day 2 upper or lower bound estimate was unreasonably large, i.e. greater than the national record accumulation amount for lowland (200m or less above sealevel) snowfall (49 inches, NOAA's National Climatic Data Center, 2019). Twenty-two participants were excluded as a result of this procedure, leaving 398 participants in the following analyses.

1) Trust.

i. Hypotheses. The primary hypotheses for this research centered around whether trust was impacted by inconsistency between two consecutive forecasts for the same event, inaccuracy of the most recent forecast (when compared to the outcome), access to probabilistic forecasts, or interactions among these variables. In particular, we hypothesized that:

i. Inaccuracy would reduce trust in forecasts.
ii. Inconsistency would reduce trust in forecasts.
iii. Probabilistic forecasts would increase trust in forecasts compared to deterministic forecasts.
iv. The negative effect of forecast inaccuracy on trust would be attenuated by the inclusion of a probabilistic forecast.
v. The negative effect of forecast inconsistency on trust would be attenuated by the inclusion of a probabilistic forecast.

ii. Data Analysis Plan. Effects on trust, an ordinal variable, were analyzed with a series of Generalized Estimating Equations (GEEs) using cumulative link proportional odds regression models (see Appendix B for model details). To conduct these analyses, we used the 'multgee' package (Touloumis, 2015) for R. We specified an 'independence' working correlation structure[9] and robust standard errors to build in resistance to possible misspecifications of the working correlation structure.[10] For this and all subsequent analyses, an alpha level of .05 was used to determine statistical significance.

[9] This is a simplifying assumption that responses nested within a participant are independent of one another.
[10] A working correlation structure does not need to be specified correctly because robust standard errors, with wider confidence intervals than naïve standard errors, are agnostic to the structure specified. Therefore, even if the working correlation structure is mis-specified, the model will still generate appropriate estimates.

iii. Trust Ratings. Contrary to our predictions, inconsistency (mismatch between Forecast 1 and Forecast 2) appeared to slightly increase (rather than decrease) trust. The estimated association between inconsistency and trust ratings was significantly positive such that a 1 inch increase in the difference between Forecast 1 and 2, compared to otherwise equivalent trials (inaccuracy and format held constant), *decreased* the odds of trust reduction (i.e. increased trust) by approximately 8%, *estimated odds ratio = 0.93, 95%CI = (0.92, 0.94), p<.001*.

Meanwhile, inaccuracy, the degree of mismatch between Forecast 2 and the outcome appeared to decrease trust as predicted. The estimated association between forecast inaccuracy and trust was significantly negative such that a 1 inch difference between Forecast 2 and the observed accumulation, compared to otherwise equivalent trials (inconsistency and format held constant), *increased* the odds of trust reduction (reduced trust) by approximately 15%, *estimated odds ratio = 1.15, 95%CI = (1.14, 1.17), p<.001*. To reiterate, although the effect of inaccuracy on trust confirmed our hypothesis, the effect of inconsistency did not. Inaccuracy had a negative association with trust (decreased trust) whereas inconsistency had a slight positive association with trust (increased trust).

Previous research suggested that inconsistent forecasts had a smaller effect on trust when forecasts were accurate (Burgeno & Joslyn, 2020). To better understand the relationship between inconsistency, inaccuracy and trust with naturalistic forecast data, we conducted exploratory analyses with inaccuracy dichotomized at 3 inches (roughly the mean of inaccuracies). Indeed, the strength of the positive association between inconsistency and trust differed significantly across levels of forecast accuracy, *p<.001* such that it was weaker for trials with higher forecast accuracy (within 3 inches of the observed outcome), *estimated odds ratio=0.95, 95%CI= (0.94,0.96)*, compared to equivalent trials (forecast format held constant)

with lower accuracy (more than 3 inches from the outcome), *estimated odds ratio=0.87, 95%CI= (0.86,0.88)*. This suggests that as with previous research accuracy attenuated the association between inconsistency and trust, although here, the effect of inconsistency is positive rather than negative. In other words, higher accuracy tended to reduce differences in trust due to consistency.

As hypothesized, probabilistic forecasts increased trust. The estimated association between forecast format and trust was significant such that when trials included probabilistic forecasts, compared to equivalent trials with deterministic forecasts (inaccuracy and inconsistency held constant), the odds of reduced trust decreased (trust increased) by approximately 20%, *estimated odds ratio= 0.80, 95%CI = (0.65,0.99), p =.04*.

The association between inconsistency and trust also differed significantly across forecast format, *p<.001*. In particular, the positive association between inconsistency and trust was stronger (farther from odds ratio=1) for trials that included probabilistic forecasts, *estimated odds ratio=0.90, 95%CI= (0.88,0.91)*, compared to equivalent trials (inaccuracy held constant) with deterministic forecasts, *estimated odds ratio=0.96, 95%CI= (0.94,0.97)*. In other words, probabilistic forecasts were associated with a stronger increase in trust due to inconsistency compared to equivalent trials with deterministic forecasts (for the general pattern, see Figure 1, Panel A).

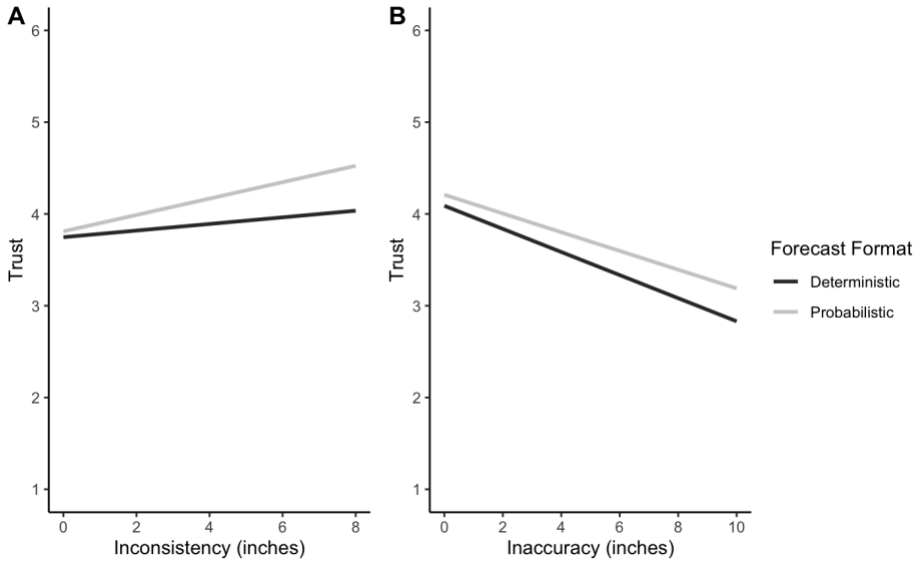

Figure 1. Panel A. Trust Rating by Inconsistency and Forecast Format. Panel B. Trust Rating by Inaccuracy and Forecast Format. Note that unlike the analyses reported above, these figures do not control for the effect of other variables. They merely provide an illustration of the general pattern.

Similarly, in support of our hypothesis, the association between inaccuracy and trust differed significantly by format, *p=.04*. The negative association was weaker for trials that included probabilistic forecasts, *estimated odds ratio=1.14, 95%CI= (1.12,1.15)*, compared to equivalent trials (consistency held constant) with deterministic forecasts, *estimated odds ratio=1.17, 95%CI = (1.14, 1.19)*. In other words, as hypothesized, probabilistic forecasts attenuated the negative effect of inaccuracy on trust, compared to equivalent trials with deterministic forecasts (for the general pattern, see Figure 1, Panel B).

Taken together, these results suggest that as predicted, probabilistic forecasts increase trust and interact with the effects on trust due to both inconsistency and inaccuracy. Here, with more realistic forecasts, unlike previous experiments, inconsistency increased rather than decreased trust. Moreover, the positive association between inconsistency and trust was enhanced by probabilistic forecasts. At the same time probabilistic forecasts attenuated the reduction in trust due to inaccurate forecasts.

2) Accumulation Estimates, Ranges and Closure Decisions.

Next we examined participants' snow accumulation estimates and closure decisions. We hypothesized that:

vi. The most recent forecast (Forecast 2) would have a greater impact on participants' outcome estimates requested after Forecast 2 was shown, than would the initial forecast, (Forecast 1) because participants would intuitively understand that the most recent forecast was more accurate.

vii. Inconsistency would increase uncertainty expectations, defined here as the range of anticipated outcomes ("as much as", "as little as"). We also asked what further impact forecast format (deterministic, probabilistic) would have on uncertainty expectations.

viii. Probabilistic forecasts would enhance decision quality, defined here as the expected value of the decision (see calculation below).

i. Data Analysis Plan. The continuous variables, snow accumulation estimates, uncertainty expectations, and decision quality, were analyzed using linear mixed model regressions[11]. A t statistic (coefficient divided by its standard error) and alpha levels of .05 were used to determine whether the coefficient of each predictor variable was significantly different from 0, i.e. whether the contribution of a given predictor was significant. School closure decisions were analyzed as a binary variable, modeled with a series of binary logistic GEEs (see Appendix C for model details). To conduct these analyses, we used the 'geepack' package for R (Højsgaard, Halekoh, & Yan, 2006), with robust standard errors. We specified an 'independence' working correlation structure and binomial family.

[11] Linear mixed model regression analyses are also capable for accounting for clustered responses.

ii. Snow Accumulation Estimates. Participants were asked to estimate the amount of snow accumulation they expected to observe on Wednesday based on the Monday and Tuesday forecasts. A linear mixed model regression was conducted on snow accumulation estimates, with three continuous predictor variables (Forecast 1 value, Forecast 2 value, inconsistency) and the categorial predictor, forecast format (deterministic, probabilistic) entered simultaneously with the inconsistency by forecast format interaction.[12]

As hypothesized, the second forecast was a much better predictor of snow accumulation estimates than was the first forecast. That is, for every 1 unit increase in the second forecast, a .87 unit increase in estimated snow accumulation was predicted, $t(51330) = 322.72$, $p<.001$.[13] In contrast, for every 1 unit increase in the first forecast, only a .08 unit increase in estimated snow accumulation was predicted, $t(51330) = 30.11$, $p<.001$. In addition, there was an unpredicted effect of inconsistency. Inconsistency slightly but significantly reduced estimates. More specifically, every 1 unit increase in inconsistency predicted a .03 unit decrease in estimated snow accumulation, $t(51330) =5.91$, $p<.001$. The main effect of forecast format failed to reach significance, $t(411) = .72$, $p=.47$. However, the inconsistency by forecast format interaction was marginally significant, $t(51330)=1.97$, $p=.05$, such that the reduction in estimates due to inconsistency was stronger for deterministic forecasts than for probabilistic forecasts. In sum, as predicted, these results suggest that participants weighted the most recent forecast 10 times more heavily than the earlier forecast, in their own estimate.

iii. Range Estimates. In order to determine whether participants inferred greater uncertainty when forecasts were inconsistent, the range of anticipated outcomes was calculated

[12] Inaccuracy was not included as a predictor because participants had not experienced a trial's forecast accuracy at the point they made an estimate.
[13] Note that, due to the inclusion of random effects, R^2 is uninterpretable for mixed model regressions.

by subtracting participants' minimum from their maximum estimated number of inches that they would not be surprised by, taken after the second forecast. A wider range suggests greater perceived uncertainty. A linear mixed model regression was conducted on range of outcomes, with inconsistency, forecast format, and the inconsistency x forecast format interaction entered simultaneously as predictors. Confirming our hypothesis, forecast inconsistency tended to increase uncertainty expectations. More specifically, every 1 inch increase in inconsistency predicted a .60 inch increase in range, $t(51340)=88.23, p<.001$. The main effect of forecast format did not reach significance, $t(404)=1.39, p=.17$. However, the inconsistency by forecast format interaction was significant such that participants who received probabilistic forecasts expected a smaller range of values for lower magnitude inconsistencies, and a larger range of snow accumulation values for higher magnitude inconsistencies, compared to participants who received deterministic forecasts, $t(51340)=10.37, p<.001, B=.10$ (see Figure 2) amplifying the effect.

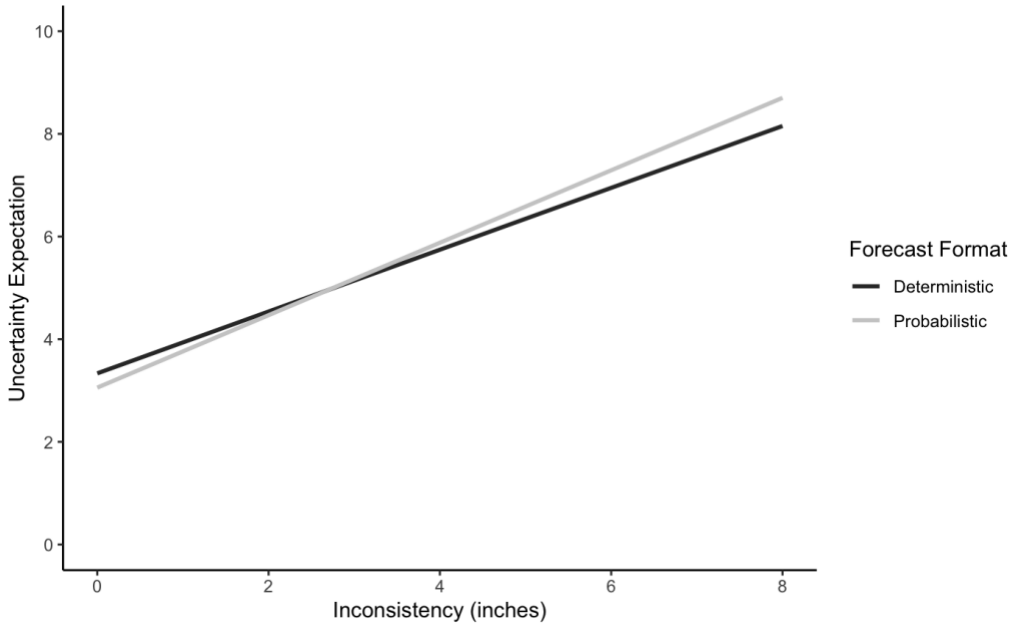

Figure 2. Uncertainty Expectations by Inconsistency and Forecast Format.

iv. Decision Quality. Next we examined decision quality defined as the expected value of the choice (Bernoulli, 1954). Because the task involved only losses (cost of closure or penalty), we refer to it as the expected cost. First, the probability weighted costs of optimal and actual decisions (on each trial) were calculated. The expected cost of keeping a school open was the product of the 6-point penalty and the chance of receiving it (the probability of observing 6 or more inches of snow accumulation associated with the second forecast). The alternative choice, to close a school, was assigned the cost of closing, 2 points. The optimal choice on any given trial was the one with the lowest cost. Next, a difference score was calculated on each trial by subtracting the expected cost of the participant's choice from the optimal choice on that trial (henceforth referred to as expected cost difference). A "0" difference indicates that the participant made the optimal choice. Otherwise the value is negative. Then, a linear mixed model regression analysis was conducted on the expected cost difference, with forecast format, inconsistency, and the inconsistency x forecast format interaction entered simultaneously as predictors[14].

Confirming our hypothesis, the expected cost difference was smaller (decision quality was better) for probabilistic compared to the deterministic forecasts. In particular, shifting from the deterministic to the probabilistic format predicted a .06 unit decrease in the expected cost difference, $t(51736)=10.10, p<.001$. Although smaller than the effect of forecast format, there was also an unpredicted increase in decision quality due to inconsistency. For every 1 unit increase in inconsistency, there was a .02 unit decrease in expected cost difference, $t(51736)=21.19, p<.001$. Additionally, the inclusion of probabilistic forecasts increased decision quality for lower magnitude forecast inconsistencies compared to deterministic forecasts, but this

[14] Inaccuracy was not included as a predictor because participants had not learned the outcome at the point they made a decision.

benefit diminished as inconsistencies increased in size. That is, the inconsistency by forecast format interaction was significant such that the probabilistic forecast reduced the expected cost difference (increased decision quality) for smaller inconsistencies, but less so for larger inconsistencies, where decision quality was already higher, *t(51736)=6.73, p<.001, B=.008* (see Figure 3).

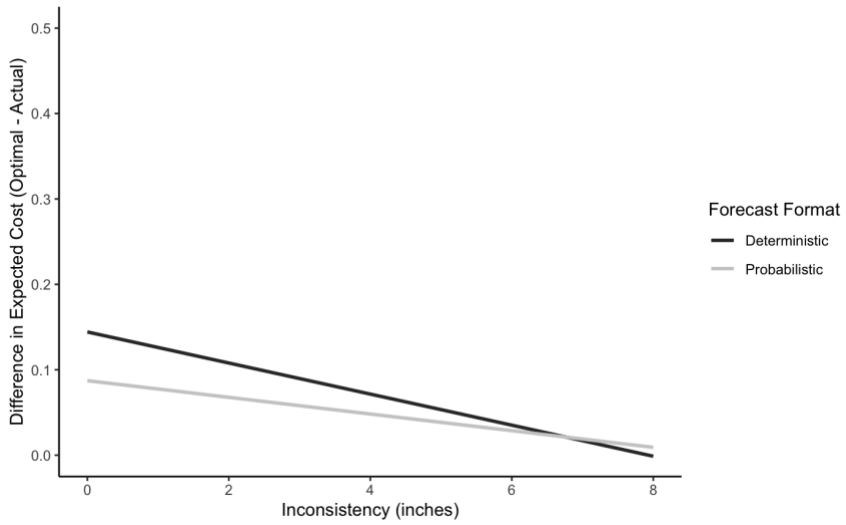

Figure 3. Expected Value Difference by Inconsistency and Forecast Format

To better understand the decision errors participants made we next examined the difference in participants decisions to close schools above and below the optimal decision threshold. According to expected utility theory, it is optimal to close schools whenever the cost to close is less than the expected value of the penalty (Murphy, 1977). The chance of 6 or more inches of snow at which this occurs can be determined by dividing the cost to close by the potential penalty of not closing (2 points /6 points = .33). Thus, the economically optimal strategy would be to close schools when the probability of 6 or more inches of snow was 33% or higher, and to keep schools open otherwise. By this standard, as is common with decisions that involve only losses (Tversky & Kahneman, 1979), the majority of decision errors (65%) were

risk-seeking (participants kept schools open when they should have closed) as opposed to risk-averse decisions (closing schools when they should stay open). Binary logistic GEE models were used to examine the associations between closure decisions (open or close) and forecast format, inconsistency, and a categorical variable, optimal decision (optimal to stay open or close), as well as forecast format by inconsistency and forecast format by optimal decision interactions. There were three models: one with all main effects entered simultaneously, and one for each of the two interaction effects (controlling for all main effects; see Appendix C).

Indeed, participants tended to follow the optimal strategy. The estimated association between optimal decision and actual closure decisions was significantly positive such that a day 2 forecast probability at or above 33% increased the odds of deciding to close by approximately 3000%, *estimated odds ratio=30.39, 95% CI= (28.30, 32.60), p<.001*.

Participants also closed more often as the inconsistency in forecasts increased. The estimated association between inconsistency and closure decision was significantly positive, such that a 1 inch difference in Forecast 1 and Forecast 2, compared to otherwise equivalent trials (forecast format and threshold orientation held constant), increased the odds of deciding to close by approximately 44%, *estimated odds ratio=1.44, 95%CI =(1.42, 1.46), p<.001*. In addition, the association between optimal decision and actual closure decisions was stronger for larger inconsistencies compared to smaller inconsistencies, *estimated odds ratio =1.67, 95% CI = (1.59,1.77), p<.001*.[15]

As shown in the expected value analysis, participants made better decisions with probabilistic forecasts. The estimated association between optimal decision and closure decisions significantly varied across forecast format, *p<.001*. In particular, probabilistic

[15] This may be explained by the fact that magnitude of inconsistency was positively correlated with forecasted probability of greater than 6 inches accumulation (r=.67, p<.001).

forecasts supported greater differentiation across the decision threshold, *estimated odds ratio = 22.9, 95% CI = (21.3, 24.5)*, compared to deterministic forecasts (inconsistency held constant), *estimated odds ratio = 11.7, 95% CI = (9.8,14.0)*. That is, probabilistic forecasts decreased the odds of deciding to close when it was optimal to keep a school open and increased the odds of deciding to close when it was optimal to close compared to those who received deterministic forecasts, coinciding with the expected value analysis.

Thus, examination of closure decisions above and below the optimal threshold (33% chance of 6 or more inches) aligned with the expected value analysis. Probabilistic forecasts allowed participants to make better decisions than did deterministic forecasts in both analyses. Participants also made better decisions when forecasts were inconsistent. This was due in part to the fact that inconsistency encouraged them to close the schools more often, an advantage in this task in which people tend to be risk seeking (majority of errors were *not* closing when closing was optimal).

c. Discussion/Conclusion

The experiment reported here is the first to demonstrate the impact of forecast inconsistency on trust and decision making using naturalistic forecast stimuli. It is also the first to demonstrate the benefits of probabilistic forecasts in this context. The results align with those reported in previous highly controlled experiments (Burgeno & Joslyn, 2020; Su, Burgeno, & Joslyn, 2021) suggesting that forecast inconsistency may not be as detrimental as is often assumed. It is important to consider the impact of forecast inconsistency in the context of forecast inaccuracy. 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. As shown previously in studies with highly controlled forecast stimuli (Burgeno & Joslyn, 2020) and here

with naturalistic forecast data, inaccuracy is much more detrimental to trust than is inconsistency.

Here, in contrast to the highly controlled studies cited above, the results suggest that, with naturalistic forecasts inconsistency may actually have a positive impact on trust. One potential explanation is that this affect is due to the forecast set used as stimuli. In this case, the inconsistent forecasts were predominantly ascending (the second forecast was for greater accumulation that the first) in which the increasing trend between forecasts was confirmed by the result (e.g., the observed accumulation was higher than the most recent forecast; trend was confirmed for 46 out of all 62 ascending trials, or 72% of ascending trials). People may have expected the trend to continue (Hohle & Teigen, 2015, 2018; Maglio & Polman, 2016) and confirmation of those expectations may have increased trust. Therefore, the slight positive effect of inconsistency on trust found in this experiment may be specific to situations in which the trend in forecasts is confirmed by the result. This is something that should be explored in future research.

An alternative more general explanation is that inconsistency may increase trust because it acts as an estimate of uncertainty. As with the prior research (Burgeno & Joslyn, 2020), the results reported here demonstrated that participants expected a larger range of outcomes with greater inconsistency, suggesting they perceived greater uncertainty in these forecasts. Moreover, as with numerous previous experiments (Joslyn & LeClerc, 2012; LeClerc & Joslyn, 2015; Grounds, LeClerc, & Joslyn, 2018), the inclusion of an explicit uncertainty estimate in the forecast (the probability of 6 or more inches of snow) increased trust in the forecast. It may be that when uncertainty was acknowledged in some way (either with and explicit uncertainty

estimate or inconsistency in forecasts), the forecast seems less "wrong" when the observed snow accumulation does not match the forecasted value.

Not only is inconsistency less deleterious to trust than inaccuracy but it may also provide the decision-maker with useful and decision-relevant information. In particular, interpreting forecast inconsistency as an indication of uncertainty, may lead users to make better decisions. Decision quality operationalized here as the difference in expected cost between the optimal and participants choices, was increased both when forecasts were inconsistent and to a greater extent with probabilistic forecasts.

Indeed, the other major contribution of this research is to demonstrate the benefits of probabilistic forecasts in the face of both forecast inconsistency and inaccuracy. In general, probabilistic forecasts preserved trust and enhanced decision quality. Decision quality operationalized both as expected value and closure decisions relative to the optimal threshold was higher with probabilistic than deterministic forecasts. Looking more closely at decision errors clarified the benefits of the probabilistic forecast. In general, participants in this experiment made more risk seeking (failure to close schools when it was economically optimal) than risk averse errors (closing schools when it was NOT economically optimal). Again, this is likely due to the fact that only losses were possible in the task (Tversky & Kahneman, 1979). People tended to prefer to take the risk that they would be penalized than to pay the point cost for closing up front. However, the optimal decision analysis revealed that those with probabilistic forecasts differentiated to a greater degree across the optimal decision threshold. When provided with the probability of six or more inches of snow, participants closed schools more often when it was economically optimal to do so (underlying forecast probability was 33% or more), and

kept schools open more often when it was economically optimal to do so (underlying forecast probability was less than 33%) compared to participants using the deterministic forecast alone.

Somewhat surprisingly, forecast inconsistency also increased decision quality slightly. However, the error analysis also revealed a key difference between the benefits of forecast inconsistency the benefits of probabilistic forecasts. Forecast inconsistency appeared to encourage greater cautiousness overall. Participants closed schools more often with greater inconsistency which increased decision quality in this task in which risk seeking errors (failing to close when it was optimal) were more prominent. The increase in cautiousness with inconsistent forecasts seen here may have been due in part to the fact that with these forecast data, greater inconsistency tended to be correlated with higher forecasted snow accumulation totals in one of the two forecasts ($r=.81$, $p<.001$). Regardless of the reason, it is safe to conclude that the effect of inconsistency on closure decisions differs from that of probabilistic forecasts which was more precise. Taken together, these findings add to the growing body of literature suggesting that people can understand and use uncertainty estimates to make better decisions (Joslyn & LeClerc, 2012; LeClerc & Joslyn, 2015; Grounds, LeClerc, & Joslyn, 2018; etc.).

We were also interested in how people integrate information from differing forecasts to form their own estimates. In line with the previous research on sequential forecasts (Burgeno & Joslyn, 2020), participants' snow accumulation estimates were influenced more strongly by second than by first forecast values. In other words, although participants did not completely disregard the first forecast, they appeared to understand that the most recent forecast should be regarded as a replacement for the first. There are at least two possible explanations for this. It may be due to extensive extra-experimental experience with real weather forecasts, leading to many, oftentimes correct, intuitions about forecasts (Morss, Demuth, Lazo, 2008; Joslyn &

Savelli, 2010; Savelli & Joslyn, 2012). However, it's important to note that our forecast stimuli were realistic in the sense that second forecasts were more accurate (22.3% accurate) than were first forecasts (14.6% accurate). Participants might have learned (explicitly or implicitly) to discount first forecasts as "mostly wrong" within the context of the experimental experience.

The main limitation of the research presented here is related to the main goal: to evaluate the effects of forecast accuracy and consistency on trust and decision making in the context of naturalistic forecasts. Allowing forecasts and outcomes to vary naturally led to a loss in internal validity. In other words, some of the effects observed here may be limited to similar forecast sets. For instance, here (and perhaps in most naturalistic situations), the degree of inconsistency was correlated with forecast values, such that higher snow accumulation forecasts were included in pairs with greater inconsistency. Thus, participants increased cautiousness with inconsistent forecasts may have been due to the perception of greater uncertainty per se, or to the fact that inconsistent forecasts often included higher snow total values. Thus, future work should test these effects with different naturalistic forecast data to verify this particular effect. Importantly, the main results reported here align with a growing body of highly controlled experimental research that suggests both a limited negative impact of forecast inconsistency and several benefits.

In sum, there is now strong and converging evidence that the effect of forecast inconsistency is not as problematic as once thought and may also confer some benefits to forecast users. When compared to consistent forecasts, inconsistent forecasts appear at least as trustworthy, if not more so, especially in the context of inaccuracy. That is likely because they signal greater uncertainty to the decision-maker, which in turn may lead to more conservative decision strategies. It is also clear and in line with previous research, that specific uncertainty

estimates preserve trust in the context of naturalistic forecasts and outcomes, especially as inaccuracy increases. Probabilistic forecasts also allow users to make better decisions from an economic perspective. Based on this converging evidence, we recommend that providers of such information should not artificially preserve consistency at a potential loss to accuracy. Updating forecasts (even at a sacrifice to consistency) and including well calibrated uncertainty estimates, can preserve trust in the information source as well as improve decisions.

References

Bernoulli, D. (1954). Exposition of a new theory on the measurement. *Econometrica*, *22*(1), 23-36.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1-3.

Burgeno, J. N., & Joslyn, S. L. (2020). The impact of weather forecast inconsistency on user trust. *Weather, Climate, and Society*, *12*(4), 679-694.

Hohle, S. M. & Teigen, K. H. (2015). Forecasting forecasts: The trend effect. *Judgment and Decision Making*, *10*(5), 416-428.

Hohle, S., & Teigen, K. (2018). When probabilities change: Perceptions and implications of trends in uncertain climate forecasts. *Journal of Risk Research*, 1-15.

Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software, 15*, 1-11.

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126.

Joslyn, S. L., & LeClerc, J. E. (2016). Climate projections and uncertainty communication. *Topics in Cognitive Science, 8*(1), 222-241.

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, *17*(2), 180-195.

Kadous, K., Mercer, M., & Thayer, J. (2009). Is there safety in numbers? The effects of forecast accuracy and forecast boldness on financial analysts' credibility with investors. *Contemporary Accounting Research*, *26*(3), 933-968.

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: mammography adherence following false-alarm test results. *Marketing Science*, *22*(3), 393-410.

Lazo, J. K., Morss, R. E., & Demuth, J. L. (2009). 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bulletin of the American Meteorological Society*, *90*(6), 785-798.

LeClerc, J., & Joslyn, S. (2015). The cry wolf effect and weather-related decision making. *Risk Analysis*, *35*(3), 385-395.

Losee, J. E., & Joslyn, S. (2018). The need to trust: How features of the forecasted weather influence forecast trust. *International Journal of Disaster Risk Reduction*, *30*, 95-104.

Maglio, S. J., & Polman, E. (2016). Revising probability estimates: Why increasing likelihood means increasing impact. *Journal of Personality and Social Psychology*, *111*(2), 141-158.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the US public. *Weather and Forecasting*, *23*(5), 974-991.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, *105*(7), 803-816.

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.

NOAA's NCDC. (2019). *30 Years of Seattle Snow Accumulation Data (1989–2019)* [Dataset retrieved December 2nd, 2019]. NOAA's National Climatic Data Center. https://www.ncdc.noaa.gov/cdo-web

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216-1226.

Ronfard, S., & Lane, J. D. (2018). Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. *Child Development*, *89*(2), 414-429.

Savelli, S., & Joslyn, S. (2012). Boater safety: Communicating weather forecast information to high-stakes end users. *Weather, Climate, and Society*, *4*(1), 7-19.

Su, C., Burgeno, J. N., & Joslyn, S. (2021). The effects of consistency among simultaneous forecasts on weather-related decisions. *Weather, Climate, and Society*, 1-30.

Touloumis (2015). R Package multgee: A generalized estimating equations solver for multinomial responses. *Journal of Statistical Software*, *64*(8), 1-14. https://www.jstatsoft.org/v64/i08/

Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291.

Wilson, L. J., & Giles, A. (2013). A new index for the verification of accuracy and timeliness of weather warnings. *Meteorological Applications*, *20*(2), 206-216.

Appendix A

Day 2 Forecasted and Observed Probabilities

Probabilistic Forecasts	Experienced Prob of 6"+	Original Data	Observed over 6"	Total Events	Binned real probabilities
0%	3.33%	2.27%	1	30	<5%
10%	10.00%	0.00%	1	10	5 to 14
20%	20.00%	7.69%	2	10	15 to 24
30%	30.00%	36.36%	3	10	25 to 34
40%	40.00%	40.00%	4	10	35 to 44
50%	50.00%	36.84%	5	10	45 to 54
60%	60%	64%	6	10	55 to 64
70%	70%	100%	7	10	65 to 74
80%	80%	100%	8	10	75 to 84
90%	90.00%	100.00%	9	10	85 to 94
100%	100.00%	100.00%	10	10	95 to 100
Total			56	130	

Appendix B

Trust analyses: GEE model descriptions by hypotheses

Statistical Hypotheses
Hypothesis 1: Is Inaccuracy associated with Trust?
Hypothesis 2: Is Inconsistency associated with Trust?
Hypothesis 3: Is Forecast Format associated with Trust?
Hypothesis 4: Does the association between Inaccuracy and Trust differ across Forecast Formats?
Hypothesis 5: Does the association between Inconsistency and Trust differ across Forecast Formats?

Models
Hypotheses 1, 2, and 3 are addressed by Model I with predictors Accuracy, Consistency, and Forecast Format.
Hypothesis 4 is addressed by Model III with predictors Accuracy, Consistency, Forecast Format, and the Accuracy by Forecast Format Interaction.
Hypothesis 5 is addressed by Model II with predictors Accuracy, Consistency, Forecast Format, and the Consistency by Forecast Format interaction.
Model IV with predictors Consistency (high accuracy), Consistency (low accuracy), Accuracy and Forecast Format was conducted to test whether the association between consistency and trust differed across accuracy.

Appendix C

Closure Decision analyses: binary GEE model descriptions by hypotheses

Statistical Hypotheses
Hypothesis 1a: Is Optimal Decision associated with Closure Decisions?
Hypothesis 1b: Does the association between Optimal Decision and Closure Decisions differ across Forecast Formats?
Hypothesis 2a: Is Consistency associated with Closure Decisions?
Hypothesis 2b: Does the association between Consistency and Closure Decisions differ across Forecast Formats?
Hypothesis 3a: Is Forecast Format associated with Closure Decisions?
Hypothesis 3b: Does the association between Forecast Format and Closure Decisions differ across Optimal Decision?

Models
Hypotheses 1a, 2a, and 3a are addressed by Model I with predictors Optimal Decision, Consistency, and Forecast Format.
Hypothesis 1b is addressed by Model II with predictors Optimal Decision, Consistency, Forecast Format, and the Optimal Decision by Forecast Format interaction.
Hypothesis 2b is addressed by Model III with predictors Optimal Decision, Consistency, Forecast Format, and the Consistency by Forecast Format interaction.
Hypothesis 3b is addressed by Model IV with predictors Optimal Decision, Consistency, Forecast Format, and the Forecast Format by Optimal Decision interaction.

Chapter 4. GENERAL DISSERTATION DISCUSSION

Trust is critical when communicating risk information to members of the public (Slovic, 1993). Without trust, people are less likely to use the information provided, putting themselves at greater risk than necessary (Earle, Siegrist & Gutscher, 2012). Risk communication is an increasingly important issue in the world in which we now live, as the number of large-scale emergencies increases. It is commonly believed that two major threats to trust in predictions (such as weather forecasts), are inaccuracy in previous predictions and inconsistency between messages for a given event. Indeed, extensive evidence supports that inaccuracy decreases trust (Joslyn & LeClerc, 2012; Kadous, Mercer, & Thayer, 2009; Kahn & Luce, 2003; etc.). Although prior to the work presented here, there was little experimental evidence on the impact of inconsistency in facts on trust, some institutions (e.g., NOAA, 2016) consider it best practice to maintain consistency. While consistency in formatting and terminology can make it easier for users to access and interpret information (Oonk et al., 2001), consistency of facts may cause problems, especially if it is at a cost to accuracy.

A primary goal of this dissertation was to establish whether inconsistency reduces user trust in sequential forecasts from the same source and to compare the effect of inconsistency on trust to that of inaccuracy on trust. To meet these ends, four experiments were conducted: Experiments 1-3 with highly controlled forecast stimuli to isolate the effects of interest, and Experiment 4 with naturalistic forecast stimuli to test the effects with real historic forecasts in which inaccuracies and inconsistencies vary naturally. In all four experiments, participants observed two sequential snow forecasts for the same target date which varied in accuracy and consistency. Then participants provided their snow accumulation estimates, made school closure decisions based on a decision rule, and finally rated their trust in the forecasts.

Recall that the forecasts and outcomes were tightly controlled in Experiments 1-3 to minimize extraneous variables. In particular, I attempted to maintain: a realistic range of forecast values for the region in which the data was collected, equal magnitudes of inaccuracies and inconsistencies, inaccuracies and inconsistencies that cross the decision threshold, and inaccuracies in which both Forecast 1 and Forecast 2 were inaccurate. It was not possible to achieve all these controls in a single experiment. Therefore, uncontrolled characteristics were traded across experiments (see Table 1) and the answers to a subset of the primary research questions are contained in the combined results of Experiments 1-3 presented in Chapter 2.

Across all three highly controlled experiments, forecast inconsistency reduced trust but not to the extent that inaccuracy did. In addition, when forecasts were inconsistent, participants weighted the most recent forecast far more than the initial forecast when making their subjective outcome estimates. This suggests participants understood that the most recent forecast should be regarded as a replacement for the initial forecast, perhaps partially explaining the reduced impact of inconsistency. Note that this is unlike simultaneous predictions from separate sources that tend to be weighted equally (Budescu & Yu, 2007), suggesting that strategies for combining information may differ due to the temporal relationship (simultaneous vs. sequential) or source (single vs. multiple sources) of the information. Inconsistency also increased uncertainty expectations (operationalized as the range between minimum and maximum estimates; in Experiments 1 and 3 where the analysis was valid) and closure decisions (making decisions more cautious). This suggests that inconsistency communicates important decision relevant information that could be especially advantageous in dangerous situations.

Although tightly controlling forecast characteristics in Experiments 1-3 ensured that the effects observed were due to inaccuracy and inconsistency (as opposed to uncontrolled forecast

characteristics, such as unrealistic forecast values), the forecasts were necessarily artificial, and all inconsistencies and inaccuracies were small (1-2 inches). Therefore, Experiment 4 presented here in Chapter 3 tested whether these effects extend to naturalistic forecasts where inaccuracies and inconsistencies varied organically. Another primary goal of Experiment 4 was to test whether providing explicit uncertainty information can maintain trust in the context of inconsistency.

Experiment 4 was like the controlled experiments with a few exceptions. Experiment 4 used real historic snow forecast and accumulation data obtained from a collaborator at NOAA and the National Weather Service. The data set consisted of forecasted snow accumulation values, threshold probabilities, and observed outcomes from locations throughout the Northeastern U.S. for a single storm event. Inaccuracies (0-10 inches) and inconsistencies (0-8 inches) varied naturally, and to a greater extent than the controlled experiments in which inaccuracies and inconsistencies were two inches at most.

In addition, in the naturalistic forecast experiment half of participants received additional probabilistic forecast information which is dissimilar to two of the three controlled experiments in which all participants received exclusively deterministic forecasts. Again, the probabilistic forecasts used in the naturalistic forecast experiment differed from those in the single controlled experiment which included the manipulation in the sense that we adjusted some of the forecasts to ensure that they were reliable. Another difference was that Experiment 4 used cumulative link generalized estimating equations (instead of ANOVAs like the first three experiments) to accommodate for continuous as opposed to binary predictor variables and to model the ordinal trust outcome as ordinal.

Despite these differences, many of the findings from tightly controlled Experiments 1-3 were replicated in Experiment 4 with naturalistic forecast data. For instance, forecast inconsistency was not as detrimental to trust as inaccuracy. Again, participants weighted the most recent forecast far more than the initial forecast when making their subjective outcome estimates. Inconsistency also increased uncertainty expectations as it had in Experiments 1 and 3 (where the analysis was valid). In other words, inconsistency appears to signal the extent of uncertainty in the situation, such that more uncertainty was expected when forecasts were inconsistent. Also like Experiments 1-3, inconsistency increased closure decisions, making decisions more cautious. This provides additional support for the notion that inconsistency communicates important decision relevant information that could be especially advantageous in dangerous situations.

While inconsistency increased participants' uncertainty expectations in both highly controlled (Experiments 1 and 3 where the analysis was valid) and naturalistic forecast experiments, there is a possible alternate explanation for this effect in Experiment 4. That is, because there were no longer any constraints on forecast values, it is possible that the higher forecast values associated with inconsistent trials may have caused participants to expect more uncertainty. However, we also see this effect in Experiments 1 and 3 where forecast values were constrained such that inconsistencies were only 1 or 2 inches, suggesting that the effect is at least partially due to the inconsistency alone.

Similarly, while inconsistency increased closure decisions in both highly controlled and naturalistic forecast experiments, forecast values offer a possible alternate explanation for the effect in Experiment 4. That is, the increase in cautiousness with inconsistent forecasts may have been due in part to the fact that greater inconsistency was correlated with higher forecast

values. However, we see this effect in Experiments 1-3 where the average forecast values for consistent and inconsistent trials were roughly equal, suggesting that the effect is at least partially due to the inconsistency alone.

Although many of the results from Experiments 1-3 with highly controlled forecasts extended to Experiment 4 with naturalistic forecasts, there were also several differences. In particular, while the general pattern of the effects of inaccuracy and inconsistency on trust were similar such that inaccuracy was more damaging, the direction of the effect of inconsistency differed. That is, in Experiments 1-3, forecast inconsistency reduced trust but not to the extent that inaccuracy did. However, in Experiment 4 with naturalistic forecasts, not only was the reduction in trust due to inaccuracy greater in magnitude than that of inconsistency but the results also suggested that inconsistency may have a positive impact on trust. One explanation for the unexpected positive effect of inconsistency on trust in Experiment 4 is that it is an artifact of the data set used. In particular, the majority of trials had ascending forecasts where the increasing trend between forecasts was confirmed by the result (e.g., the observed accumulation was higher than the most recent forecast). Alternatively, it may be that inconsistency acted as an estimate of uncertainty. Indeed, the range of inconsistencies was much larger in Experiment 4 compared to Experiments 1-3, potentially providing a better estimate of uncertainty. By acting as a signal of uncertainty, inconsistent forecasts may have increased trust because the forecasts seemed less "wrong" when the observed snow accumulation did not match the second forecast value. This highlights the importance of testing these effects with realistic forecast data, because effects may differ due to the characteristics of the forecast.

Experiment 4 was also the only experiment presented here which tested whether reliable probabilistic forecasts offer benefits for trust, expectations, and decision making. Probabilistic

forecasts enhanced trust relative to deterministic forecasts alone and yielded higher trust in the context of both inaccuracies and inconsistencies. In particular, probabilistic forecasts dampened the negative effect of inaccuracy on trust and amplified the unanticipated positive effect of inconsistency on trust. Probabilistic forecasts also amplified the positive effect of inconsistency on participants' uncertainty estimates (the range of outcome estimates participants would not be surprised by), which may at least partially explain the relative effects of forecast inaccuracy and inconsistency on trust. That is, by enhancing perceptions of uncertainty, inconsistencies implied that the forecast would not verify exactly. In light of the results from Experiment 1 in which unreliable probabilistic forecasts provided no significant benefits, these results also highlight the importance of providing reliable (as opposed to unreliable) probabilistic information.

Utilizing real historic forecast data with probabilistic forecasts also enabled analysis of decision quality because having the probabilities made it possible to calculate expected value of participants decisions (the probability weighted outcome; Bernoulli, 1954). Decision quality operationalized both as expected value and closure decisions relative to the optimal threshold was higher with probabilistic than deterministic forecasts. That is, probabilistic forecasts significantly increased expected value as has been seen in many previous studies (Joslyn & LeClerc, 2012; LeClerc & Joslyn, 2015; Grounds, LeClerc, & Joslyn, 2018; etc.). Looking more closely at decision errors clarified the benefits of the probabilistic forecast. Regardless of whether participants had probabilistic forecasts, their decisions tended to follow the economically optimal strategy. That is, they generally decided to close schools when the probability weighted cost of staying open was equal to or greater than the cost of closing and they decided to stay open when the probability weighted cost of staying open was less than the cost of closing. However, participants who received probabilistic forecasts differentiated to a

greater degree across the optimal decision threshold, helping them to perform closer to optimal than those who received only deterministic forecasts. Somewhat surprisingly, forecast inconsistency also increased decision quality slightly. Yet, the error analysis revealed a key difference between the benefits of forecast inconsistency the benefits of probabilistic forecasts. Forecast inconsistency, appeared to encourage greater cautiousness overall. Participants closed schools more often with greater inconsistency which increased decision quality in this task in which risk seeking errors (failing to close when it was optimal) were more prominent. That is, the effect of inconsistency on closure decisions differs from that of probabilistic forecasts which was more precise.

These results suggest that not only is inconsistency less detrimental to trust than previously thought, but that inconsistency may provide important, decision-relevant information about the extent of uncertainty in the situation. That is, inconsistency may serve a similar function as explicit uncertainty estimates which people can use to more effectively determine the optimal course of action. These findings add to the mounting body of literature supporting that people are capable of understanding and benefitting from more complex information than previously thought (Joslyn & LeClerc, 2012; LeClerc & Joslyn, 2015; Grounds, LeClerc, & Joslyn, 2018; etc.).

To ensure that the effects observed are not explained by the forecast characteristics of the specific historic data set used in Experiment 4, future research should retest these effects using a different set of historic forecasts. Also note that, in the experiments reported here, inconsistency referred to differences in weather outcomes rather than advice about what to do. Although this difference seems subtle, it is possible that inconsistency in advice is less well tolerated. In addition, people have a lot of intuitions about weather (Morss, Demuth, Lazo, 2008; Joslyn &

Savelli, 2010; Savelli & Joslyn, 2012) due to their extensive experience with forecasts and outcomes. In novel situations (e.g., Covid-19) people might not understand why there are inconsistencies and may be less forgiving as a result. Therefore, in the Covid-19 mask messaging example provided earlier, the results of this dissertation suggest that the WHO may have more effectively supported trust and appropriate protective action by proactively acknowledging the slightly more complicated truth, that masks effectively protect members of the general public but that we need to save masks for health care workers. Future studies should also test whether the results of this dissertation in fact generalize beyond the weather domain (e.g., to climate change, medical, financial contexts, etc.), to different time horizons, to advice, and to decisions involving different consequences.

The results of this dissertation suggest that in the context of weather, if trust is a concern, accuracy should be prioritized over consistency. In other words, if more reliable information becomes available it should be provided to users and in fact they can make use of it. For instance, inconsistency can communicate information about the extent of uncertainty in the situation, and can help users make better, more cautious decisions. In addition, the results of this dissertation also suggest that reliable uncertainty estimates should be provided if available. Explicit uncertainty estimates enhanced the benefits provided by inconsistency. This adds to the growing body of research suggesting that members of the public can understand and make good use of fairly complex information.

Final References

Budescu, D. V., & Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153-177.

Earle, T. C. (2010). Trust in risk management: A model-based review of empirical research. *Risk Analysis: An International Journal*, *30*(4), 541-574.

Earle, T. C., Siegrist, M., & Gutscher, H. (2012). Trust, risk perception and the TCC model of cooperation. In Trust in Cooperative Risk Management (pp. 19-68). Routledge.

Grounds, M. A., LeClerc, J. E., & Joslyn, S. (2018). Expressing flood likelihood: Return period versus probability. *Weather, Climate, and Society*, *10*(1), 5-17.

Jingnan, H. (2020). Why There Are so Many Different Guidelines for Face Masks for the Public. *National Public Radio*. Available online at: https://www.npr.org/sections/goatsandsoda/2020/04/10/829890635/why-there-so-many-different-guidelines-for-face-masks-for-the-public (accessed May 3rd, 2022)

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126.

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, *17*(2), 180-195.

Kadous, K., Mercer, M., & Thayer, J. (2009). Is there safety in numbers? The effects of forecast accuracy and forecast boldness on financial analysts' credibility with investors. *Contemporary Accounting Research*, *26*(3), 933-968.

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: mammography adherence following false-alarm test results. *Marketing Science*, *22*(3), 393-410.

Lazo, J. K., Morss, R. E., & Demuth, J. L. (2009). 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bulletin of the American Meteorological Society*, *90*(6), 785-798.

LeClerc, J., & Joslyn, S. (2015). The cry wolf effect and weather-related decision making. *Risk Analysis*, *35*(3), 385-395.

Løhre, E., Sobkow, A., Hohle, S. M., & Teigen, K. H. (2019). Framing experts'(dis) agreements about uncertain environmental events. *Journal of Behavioral Decision Making*, *32*(5), 564-578.

Losee, J. E., & Joslyn, S. (2018). The need to trust: how features of the forecasted weather influence forecast trust. *International journal of disaster risk reduction*, *30*, 95-104.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the US public. *Weather and Forecasting*, *23*(5), 974-991.

National Oceanographic and Atmospheric Administration. (2016). *Risk communication and behavior: Best practices and research findings*. Silver Spring, MD : U.S.

NRC. (2006). Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts. National Research Council, Washington, D.C., 124 pp.

Oonk, H. M., Smallman, H. S., & Moore, R. A. (2001). Evaluating the usage, utility and usability of web-technologies to facilitate knowledge sharing. In Proceedings of the Command and Control Research & Technology Symposium.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216.

Ronfard, S., & Lane, J. D. (2018). Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. *Child Development*, *89*(2), 414-429.

Savelli, S., & Joslyn, S. (2012). Boater safety: Communicating weather forecast information to high-stakes end users. *Weather, Climate, and Society*, *4*(1), 7-19.

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis*, *13*(6), 675-682.

Su, C., Burgeno, J. N., & Joslyn, S. (2021). The effects of consistency among simultaneous forecasts on weather-related decisions. *Weather, Climate, and Society*, 1-30.

Twyman, M., Harvey, N., & Harries, C. (2008). Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making*, *3*(1), 111.

WHO. (2020). Listing of WHOs response to Covid-19. World Health Organization.

Wilson, L. J., & Giles, A. (2013). A new index for the verification of accuracy and timeliness of weather warnings. *Meteorological Applications*, *20*(2), 206-216.