The Impact of Forecast Inconsistency On User Trust

Jessica Noel Burgeno

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Susan Joslyn

Andrea Stocco

Program Authorized to offer degree:

Psychology

University of Washington

Abstract

The Impact of Forecast Inconsistency On User Trust

Jessica Noel Burgeno

Chair of Supervisory Committee:

Susan Joslyn

Department of Psychology

For high impact weather events, forecasts often start days in advance. Meteorologists believe that consistency among subsequent forecasts is important to user trust and can therefore 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 on user trust although the reduction in trust due to inaccuracy is well documented (Gupta, Bisantz, & Singh, 2001; Kahn & Luce, 2003; Joslyn & LeClerc, 2012). 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 suggesting that it is inadvisable for forecasters to sacrifice accuracy in favor of consistency.

The Impact of Forecast Inconsistency on User Trust

Forecasts for major weather events often begin days in advance. The weather models, upon which public 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 to update the forecast they provide to the public. Sometimes they are reluctant to do so out of fear that inconsistency in subsequent forecasts will be confusing and negatively affect user trust. Indeed, the maintenance of forecast consistency is considered best practice for some institutions, including the National Oceanic and Atmospheric Administration (NOAA, 2016). However, because accuracy generally increases as lead times decrease, the choice to maintain consistency 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 the chance of an extreme event was decreasing. However, forecasters failed to 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. 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.

Surprisingly, there is little experimental research that investigates the effect of inconsistency on trust, although there is some research on other, potentially related issues. For instance, people are more inclined to trust news that is congruent with their previous attitudes than news that is incongruent (Eiser & White, 2005). In addition, people have higher confidence in decisions made based on advice from financial advisors who agree with one another (Budescu et al., 2003). Nonetheless, when presented with inconsistent financial advice from multiple advisors simultaneously, participants own estimates appeared to be due to a simple averaging strategy, suggesting that all of the advice is considered and weighted equally (Budescu & Yu, 2007). It is important to note however, that the weather situation discussed above is different in that it involves sequential forecasts from the same source. The only work of which we are aware that addressed this exact situation was a survey conducted among African Americans who chose not to evacuate leading up to Hurricane Katrina. The survey found that inconsistency in evacuation orders attenuated perceptions of the severity of the event and influenced evacuation decisions (Elder et al., 2007).

Although there is little or no experimental evidence for the impact of inconsistency of sequential forecasts on trust, there is extensive support for the negative effect of inaccuracy on trust. For example, in a virtual driving simulation with icy road conditions participants trusted weather warnings less when alarms were activated at a lower relative to a higher threshold, presumably due to the reduction in false alarms (Gupta, Bisantz, & Singh, 2001). 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). Participants assigned a road salting task based on overnight low temperature forecasts reported significantly

higher trust in low as compared to high- error forecasts and were more likely to take protective action (Joslyn & LeClerc, 2012).

Therefore, because the maintenance of consistency could be at the sacrifice to accuracy in situations in which more recent information is more accurate, determining the impact of inconsistency is critical. The purpose of the research reported here is to determine whether inconsistency in sequential forecasts from the same source reduces user trust, and if so, how that relates to the reduction in trust due to inaccuracy. To answer these questions, we conducted three lab-based experiments in which participants were asked to take the role of a decision consultant responsible for advising schools when 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.

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.

Method

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.

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

Procedure. 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 and cost/loss structure. Participants were tasked with providing decision advice to schools regarding whether they should stay open or close due to an upcoming snow-storm. 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 their goal was to retain as many points as possible by giving their best advice. Points could be spent at a rate of 2 per school closure recommendation to reflect the cost of makeup days. There was no cost for recommending that a school stay open; however, if participants recommended staying open and six or more inches of snow accumulation was observed, a 6-point penalty was deducted from their score to reflect the risk of travel in dangerous road conditions. 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). Forecasts were provided by one of four fictitious forecast providers, TruWeather™, Weather Now™, Weather Direct™, and AccuCast™, each making snow accumulation forecasts for a block of six consecutive trials. 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.

The two snow accumulation forecasts for Wednesday were presented sequentially, centered on separate screens. The current date and day 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, 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 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 in the lower right-hand corner of the screen to advance to the next screen. Once the next button was pressed, they could not go back and change answers on the previous screen. After the second forecast, participants were shown a decision screen. The current date and day (Tuesday) was shown in bold font in the upper left-hand corner of the screen. Because, in reality decision makers would likely have better memory for the current than for the previous days' forecast, a reminder of the Tuesday forecast for Wednesday was also provided in a box in the upper right-hand corner of the decision screen. In the middle of the screen were two buttons labeled "close" or "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." 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 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: Monday forecast for Wednesday, Tuesday forecast for Wednesday, Tuesday night school closure decision, and Wednesday outcome. Participants performed four practice trials before the test trials began.

Stimuli. There were four blocks of six trials with four experimental and two filler trials per block, for a total of 24 trials, 16 experimental and eight 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 more extreme values would be unrealistic and/or would make the most advantageous closure decision obvious.

Half of the 4 experimental trials within each block were accurate and half were inaccurate. Accuracy was defined as an exact match between the Tuesday forecast and the accumulation observed on Wednesday. All inaccurate experimental trials were inaccurate by a 2 inch difference between the second (Tuesday) forecast and the Wednesday 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 less inaccurate in the sense that the forecast would suggest the correct response. Half of accurate trials in each block were correct rejections (CR), in which both the second forecast and the observed accumulation values were below the 6 inch

decision threshold. Half of accurate trials in each block were hits, in which the second forecast and the observed accumulation values were above the threshold. Half of inaccurate trials in each block were false alarms (FA), in which the second forecast was at or above the 6 inch decision threshold and the observed accumulation was below the threshold. Half of inaccurate trials in each block were misses, in which the second forecast was below the 6 inch decision threshold and the observed accumulation was at or above the threshold.

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 second forecast and observation values 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).

Half of experimental trials were consistent and half were inconsistent. 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, ensuring that inaccurate trials crossed the 6 inch decision threshold, and that inaccurate inconsistent trials remained inaccurate for both first and second forecasts. Therefore, in Experiment 1 the inconsistencies in inaccurate inconsistent trials were 1 inch while the accurate inconsistent trials were 2 inches (we return to this issue in the discussion). Out of concern that the effect of consistency may be small, each block of trials contained exclusively consistent or inconsistent trials to allow for a build up of trust or distrust over several trials. 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). Thus, there were four different types of experimental trials: accurate consistent, accurate inconsistent, inaccurate consistent, and inaccurate inconsistent. Each trial type accounted for a quarter of the 16 experimental trials.

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"). In contrast, 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. We found no significant main effect or attenuating effects of forecast format, perhaps because the forecasts were not in fact reliable. As a result, the conditions were combined and this manipulation has been omitted from subsequent sections of this paper.

Design. A 2(accuracy) by 2(consistency) within groups design was used. Accuracy had two levels, accurate and inaccurate. Consistency also had two levels, consistent and inconsistent. All analyses reported below combine conditions in the between groups forecast factor (probabilistic/deterministic).

Results

The primary question for this research was whether forecast inconsistency caused a reduction in user trust and how that compared to the loss of trust due to inaccuracy. Where appropriate, Cohen's d and partial eta squared are provided to allow 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. We excluded the five participants who reported higher average minimum than maximum snow accumulation estimates leaving a total of 363 participants.

The first set of analyses revealed that inconsistency did in fact significantly reduce trust but not to the extent that inaccuracy did. In order to determine the impact of inaccuracy and inconsistency on trust, 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 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), *F(1, 361)=17.35, p<.001, Cohen's d=.43* (see Figure 1). In addition, 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), *F(1, 361)=124.83, p<.001, Cohen's d=1.18*. Notice that the magnitude of the effect of inaccuracy appears to be substantially larger than that of inconsistency. Additionally, there was a 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=.31*.

It is also worthwhile to compare the options meteorologists might often face: 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 trials (M=3.25, SD=0.91) than in inaccurate consistent trials (M=2.96, SD=1.01), *t(362)=5.72, p<.001*. This analysis is especially interesting 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. Indeed, participants rated consistent trials significantly more trustworthy (M= 3.19, SD=0.79) than inconsistent trials (M= 3.05, SD=0.82), $t(362) = 4.51$, $p < .001$,but the effect size was small $Cohen's\ d = .17$.

We next investigated how people incorporated the information in inconsistent forecasts in 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, $F(2, 2891) = 421.89$, $p < .001$, $R^2 = .23$.

Clearly receiving two conflicting forecasts had an effect on trust and expected outcomes. In order to determine whether consistency also effected decisions to close schools, a 2 (consistency: consistent, inconsistent) by 2 (threshold orientation: below, above) repeated measures ANOVA was conducted on school closure decisions. Here we included only the most

extreme second forecast values of 4 inches (below threshold) and 7 inches (above threshold). It is important to note that within each level of consistency, the second forecast values were identical and each value (4 and 7 inches) appeared twice. Not surprisingly, participants closed significantly more often when 7 inches of snow was forecasted (M= .82, SD=.17) than when 4 inches was forecasted (M= .09, SD=.17), $F\ (1, 362) = 91.79, p<.001, \eta^2 = .20$, suggesting that they understood the task and adopted the goal we assigned them. In addition, participants closed slightly but significantly more often on consistent trials (M= .51, SD= .13) than inconsistent trials (M= .40, SD= .20), $F\ (1, 362)= 3820.45\ p<.001, \eta^2 = .91$. The consistency by second forecast value interaction was also significant suggesting that there is a greater difference between consistent (M = .96, SD= 0.16) and inconsistent (M = .69, SD= 0.29) trials above the decision threshold (7 inches) than below the threshold (consistent, M= 0.07, SD= 0.19; inconsistent, M= 0.11, SD= 0.24), $F(1, 362)= 201.35, p<.001, \eta^2 = .36$ (see Figure 2). More specifically, above the threshold, more people chose to close when forecasts were consistent than when they were inconsistent.

This difference in decision strategy may have been due to a perception of greater uncertainty when forecasts were inconsistent. 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=3.38, SD=1.73) than when forecasts were consistent (M=3.27, SD=1.56), $t(362)=2.11, p=.036$. It is important to note that consistent and inconsistent trials 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, the average trust in consistent trials would increase over trials within a block (positive correlation) and the average trust in inconsistent trials would decrease (negative correlation) over trials within a block. With the exception of one block (the first block for participants who received an inconsistent block first, $r = -.98, p<.00625$), none of the correlations between trust and trial number reached significance.

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. This may have minimized the difference in trust between inaccurate consistent and inaccurate inconsistent trials. Experiment 2 was conducted to address this issue.

However, it is important to note that the crucial comparison that forecasters most likely face was unaffected by this issue. There was significantly greater trust in accurate inconsistent forecasts than 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 consistency 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. This suggests that as far as consumer trust is concerned, meteorologists may be better served updating forecasts when they believe that better information is available.

In addition, these results suggest that when forecasts are inconsistent, they are not weighted 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. Alternatively, the second forecast could simply be better remembered, as a reminder was provided. This issue will also be addressed in Experiment 2.

In addition to the impact on trust, inconsistency appears to have promoted less conservative decision strategies. Participants chose to close schools less often overall when forecasts were inconsistent, and especially for forecasts above the decision threshold. This may have been due to the perception of greater uncertainty when forecasts were inconsistent suggested by the fact that the greater ranges of outcomes were expected with inconsistent forecasts relative to consistent forecasts. Here, in a loss scenario, greater uncertainty may have promoted greater risk seeking (Kahneman & Tversky, 1979).

However it is important to note that the results from the primary analyses in this study should be interpreted with caution. In an effort to control for a number of extraneous but potentially influential variables, two confounds remained. First, the magnitudes of all inaccuracies were 2 inches, while the inconsistencies in inaccurate inconsistent trials were only 1 inch. Second, all inaccuracies crossed the 6 inch decision threshold while inconsistencies

crossed the threshold in only half of inconsistent trials (the accurate ones). Experiment 2 was conducted to address the first confound.

Experiment 2

In Experiment 2, the range of forecast values was expanded so 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 emphasized second forecast values contributing to their greater weighting in participants' own estimates. Therefore, in Experiment 2 we also manipulated the reminder to test its impact.

Method

Participants. A total of 164 University of Washington psychology students (49.1% female, mean age = 19.71 years) participated for course credit and the opportunity to earn a cash bonus.

Procedure. The computer-administered procedure was identical to that used for Experiment 1.

Stimuli. The stimuli used in this task were identical to Experiment 1 with three exceptions (see Table 2). First, Experiment 2 participants received only single-value, deterministic forecasts (recall that probabilistic forecasts were included for half the sample in Experiment 1, although these conditions were combined due to the lack of effect). Second, in Experiment 2, the range of forecasted snow accumulation values for experimental trials was greater (2-9 instead of 4-7 inches) to allow inaccurate inconsistent trials to be inconsistent by 2 inches, 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, in order to determine whether including a second forecast reminder at the time of decision contributed to an overweighting of the second forecast, half of participants received reminders on the decision screen and half did not.

Design. We used a 2(accuracy) by 2(consistency) by 2(forecast reminder) mixed model design. Accuracy and consistency were both within-subjects variables with 2 levels each, accurate and inaccurate, and consistent and inconsistent respectively. Forecast reminder was a between-subjects variable with 2 levels, present and absent.

Results

In Experiment 2, 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) by 2 (forecast reminder: present, absent) mixed model ANOVA conducted on post outcome trust indicated that 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)=12.59, p=.001, Cohen's d=.56* (see Figure 3). In addition, 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)=52.18, p<.001, Cohen's d=1.14*. Again, the effect of inaccuracy was much greater than that of inconsistency. However, although in the expected direction, the accuracy by consistency interaction did not reach significance, *F(1,160)=1.94, p=.17, Cohen's d=.22*. Moreover, trust ratings did not differ significantly for forecast reminder (M=3.52, SD=.82) compared to no reminder (M=3.27, SD= .69), *F(1,160)=3.33, p=.07, Cohen's d=.29*. Nor did any other interactions reach significance.

As with Experiment 1, a paired samples t-test indicated again that trust ratings for accurate inconsistent trials (M = 3.29, SD=0.07) were significantly higher than inaccurate consistent trials (M=3.10, SD=0.08), t(161)=2.35, p=.02. Again, blocking consistency had no effect.

In addition, pre-outcome trust ratings were significantly higher for consistent (M= 3.28, SD=0.81) than inconsistent trials (M= 3.09, SD=0.74), *t (160) = 3.52, p= .001, Cohen's d=0.24.* Although, as with Experiment 1, the pre outcome effect of consistency was smaller than the post outcome effect of consistency ($Cohen's\ d = .56$), and smaller than the large effect of accuracy on post outcome trust ($Cohen's\ d = 1.14$).

To evaluate whether including the second forecast reminder contributed to an overweighting of the second forecast when forecasts were inconsistent, a multiple regression analysis was conducted with first and second forecast values and forecast reminder entered simultaneously as predictors of snow accumulation estimates. Similar to Experiment 1, the second forecast had a much bigger impact. Additionally, it did not seem to be due to the reminder. A one unit increase in the second forecast value predicted a 0.80 unit increase in snow accumulation estimates, $\beta = .71$, $p < .001$, while a one unit increase in the first forecast value only predicted a 0.04 unit increase in snow accumulation estimates, $\beta = .08$, $p < .001$. However, forecast reminder was not a reliable predictor of snow accumulation estimates, $\beta = -.01$, $p = .76$. Overall, the three predictor model accounted for 56% of the variance in snow accumulation estimates, $F(3, 1292) = 549.89$, $p < .001$, $R^2 = .56$.

School closure decisions yielded somewhat different results than in Experiment 1. In Experiment 2 a 2 (consistency: consistent, inconsistent) by 2 (threshold orientation: below, above) repeated measures ANOVA on school closure decisions revealed that participants closed significantly more often on *inconsistent* (M= .61, SD=.48) than on consistent

trials (M= .59, SD=.49), *F (1, 160)= 9.83, p<.01,* $\eta^2 = .06$, suggesting a more conservative strategy when forecasts were inconsistent. The consistency by second forecast interaction was significant revealing a greater difference between consistent (M= 0.09, SD= 0.20) and inconsistent (M= 0.23, SD= 0.33) trials below the decision threshold (4 inches) than above the threshold (consistent, M = 0.95, SD= 0.16; inconsistent, M= 0.91, SD= 0.21), *F(1, 160)= 27.60, p<.001,* $\eta^2 = .15$ (see Figure 4). As with Experiment 1, participants closed significantly more often when 7 inches of snow were forecasted (M=.93 , SD=.25) than when 4 inches of snow were forecasted (M=.16 , SD=.37), *F (1, 160) = 1528.52, p<.001,* $\eta^2 = .91$. Thus, unlike Experiment 1, inconsistency in Experiment 2 led to a more conservative decision strategy.

To examine whether inconsistency influenced participants uncertainty perceptions, a paired samples t-test was conducted on range of expected snow accumulation. As with Experiment 1, participants estimated a significantly larger range when forecasts were inconsistent (M=3.47, SD=1.46) than when forecasts were consistent (M=3.31, SD=1.46), t(160)=2.23, p=.03. Unfortunately, in order to address the confounds present in Experiment 1 whilst controlling for all other variables, the forecast values in the consistent and inconsistent trials were not identical as they had been in Experiment 1. Instead, more extreme values (2,3, 8 and 9 inches) were used for inaccurate inconsistent first forecasts while all other values were identical to those used in Experiment 1. Consequently, it is possible that the difference in estimated range is due to differing values.

Discussion Experiment 2

The negative effects of inconsistency and inaccuracy on user trust found in Experiment 1 were replicated in Experiment 2. Again, comparing the effect sizes of inconsistency and inaccuracy, the magnitude of the effect of inaccuracy appears to be substantially larger than that

of inconsistency. Again, there was no effect of blocking consistency on trust. The consistency by accuracy interaction in Experiment 1 did not reach significance in Experiment 2, although it was in the expected direction. However, in the crucial comparison between the options meteorologists most often face, as in Experiment 1 accurate but inconsistent forecasts were trusted significantly more than inaccurate consistent forecasts. In addition, as with Experiment 1, without knowledge of accuracy, the effect of consistency on pre-outcome trust was relatively small. Thus the recommendation stands: as far as consumer trust is concerned, meteorologists are better served updating their forecasts for the sake of accuracy.

As with Experiment 1, participants' accumulation estimates were predicted more strongly by second forecast values than first forecast values. In addition, Experiment 2 revealed that the second forecast reminder was not a significant factor, ruling it out as a possible explanation for the asymmetric weighting of the two forecasts. This suggests that users understand that the second forecast should be regarded as a replacement for the first forecast and is likely to be more accurate.

As with Experiment 1, inconsistency appears to lead to the perception of greater uncertainty in terms of larger expectation ranges. However, here it promoted more conservative decision strategies, participants closed schools more often when forecasts were inconsistent especially when forecasts are below the decision threshold. In Experiment 1, inconsistency had lead to slightly less conservative strategy. This difference may be due in part to the fact that inconsistencies were larger in magnitude in Experiment 2 (all 2 inches) relative to Experiment 1 (1.5 inches on average).

All-in-all, Experiment 2 results confirm the main results of Experiment 1 suggesting that inconsistency reduces trust, although not to the degree of inaccuracy, and leads to perceptions of

greater uncertainty. 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 the larger negative effect of inaccuracy on trust relative to that of inconsistency.

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 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.

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. The second forecast reminder was manipulated once again to verify its lack of impact.

Method

Participants. A total of 160 University of Washington psychology students (50.6% female, Mean age = 19.9 years) participated for course credit and the opportunity to earn a cash bonus.

Procedure. The computer-administered procedure was identical to that used for Experiments 1 and 2.

Stimuli. The stimuli in Experiment 3 were identical to Experiment 1 stimuli 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., a forecast of 4 inches of snow accumulation on the first forecast, a forecast of 6 inches of snow accumulation on the second forecast, and an observed accumulation of 4 inches). Thus the magnitudes of all inaccuracies and inconsistencies were equal (See Table 1) and all crossed the decision threshold. However inaccuracy (defined as a match between second forecast values and the outcome) in the inaccurate inconsistent condition may have been regarded as less inaccurate by participants, because the first forecast matched the outcome.

Design. We used an identical 2(accuracy) by 2(consistency) by 2(forecast reminder) mixed model design to that used in Experiment 2.

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) by 2 (forecast reminder: present, absent) mixed model ANOVA was conducted on mean post-outcome trust. Consistent trials (M= 3.40, SD= 0.87) were rated significantly higher than inconsistent trials (M= 3.05, SD= 0.83), $F\ (1, 156) = 42.41, p<.001,$ *Cohen's d=1.04* and accurate trials (M= 3.45, SD= 0.80) were rated significantly higher than inaccurate trials (M= 3.00, SD= 0.90), $F\ (1, 156)= 66.65, p<.001,$ *Cohen's d=1.31* (see Figure 5). The effect of accuracy was again greater than that of consistency. 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 trials (M=3.21, SD=0.88) when the forecast was

accurate (consistent, M=3.70, SD=0.91; inconsistent, M=3.21 , SD=0.07) than inaccurate (consistent, M=3.11, SD=1.08; inconsistent, M=2.89, SD=0.91), *F(1, 156)= 13.72, p<.001, Cohen's d=.59*. There was no significant effect of the forecast reminder on trust, *F (1, 156) = 0.63, p=.43,* nor were any interactions significant with forecast reminder. Although trust ratings for accurate inconsistent trials (M= 3.21, SD=0.88) were higher than for inaccurate consistent trials (M=3.11, SD=1.08), unlike the first two experiments, the difference did not reach significance, *t(157) = 1.17, p=.24, Cohen's d= .10*. Again the effect of pre-outcome trust was small. A paired samples t-test revealed significantly higher 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.* Again, blocking for consistency had no impact on trust. Thus the main findings showing a greater impact of accuracy, as compared to consistency, 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 and forecast reminder, entered simultaneously as predictors revealed that the second forecast had a much bigger impact, and similar to Experiment 2, it did not seem to be due to the reminder. A one unit increase in the second forecast value predicted a .77 unit increase in estimated snow accumulation, $\beta = .80, p < .001$, while a one unit increase in the first forecast value predicted a .09 unit increase in estimated snow accumulation, $\beta = .09, p < .001$. Again, forecast reminder was not a reliable predictor of snow accumulation estimates, β=-*.05, p=.18*. Overall, the three predictor model explained a significant proportion of the variance in snow accumulation estimates, *F (3, 1260) = 535.31, p<.001,* R^2*= .56*.

The analysis of school closure decisions replicated the results of Experiment 2, showing more conservative decisions for inconsistent forecasts. A 2 (consistency: consistent, inconsistent)

by 2 (threshold orientation: below, above) mixed model ANOVA revealed that participants closed significantly more often when forecasts were inconsistent (M= 0.60 , SD= 0.21) than consistent (M= 0.57, SD= 0.16) *F (1, 157)= 19.69, p<.001,* $\eta^2 = .1$. Participants closed more when 7 inches of snow were forecasted (M= 0.95, SD= 0.09) than when 4 inches of snow were forecasted (M= 0.13, SD= 0.16) , *F (1, 157) = 1586.52, p<.001,* $\eta^2 = .91$. The consistency by second forecast value interaction was significant indicating that the difference between consistent (M= 0.05, SD= 0.16) and inconsistent (M= 0.25, SD= 0.35) forecasts was greater for trials below the decision threshold (4 inches) compared to above (7 inches) (consistent, M = 0.94, SD= 0.18; inconsistent, M= 0.91, SD= 0.24), *F(1, 157)= 52.65, p<.001,* $\eta^2 = .25$ (see Figure 6). This suggests that inconsistencies encourage a conservative decision strategy especially at low values.

As with the previous two experiments, participants estimated a significantly greater range of snow accumulation when forecasts were inconsistent (M= 3.83, SD=3.46) than consistent (M=3.38, SD=2.57), *t (157) = 3.83, p< .001, Cohen's d = .15*. However, unlike Experiment 1, consistent and inconsistent trials did not use the same forecast values the same number of times so we cannot be confident that this difference in estimated range is due to consistency alone in this experiment, rather than values that may seem more or less plausible.

Discussion Experiment 3

Experiment 3 replicated nearly all of the effects observed in the previous two experiments, with a different stimuli set designed to further address the confounds identified in the previous experiments. The negative effects of inconsistency and inaccuracy on user trust found in Experiments 1 and 2 were replicated in Experiment 3. Again, the magnitude of the effect of inaccuracy on trust was larger than that of inconsistency but by a smaller margin than in

Experiments 1 and 2. Comparing mean trust ratings for all experiments (see Table 2) reveals that the difference between consistent and inconsistent trials was amplified in Experiment 3 relative to the previous experiments. This may be because, unlike the previous experiments, all inconsistencies crossed the 6-inch decision threshold in Experiment 3.

The consistency by accuracy interaction found in Experiment 1 reemerged here suggesting that inconsistency matters more when forecasts are accurate. This is particularly striking in light of the fact that half of the inaccurate trials included first forecasts that matched the outcome (i.e. could have been considered accurate). Moreover, the effect of consistency on pre-outcome trust remained relatively small, as in the previous two experiments. Again, there is no evidence that the effect of consistency was building over the course of a block. Only the difference in trust between the accurate inconsistent forecasts and inaccurate consistent forecasts failed to reach significance here, although it was in the expected direction. Taken together, these results contribute to the building evidence for greater importance of accuracy over consistency in preserving user trust.

As in Experiments 1 and 2, in Experiment 3 accumulation estimates were influenced more strongly by the second forecast values and were not influenced by the reminder. In other words, users appear to understand that the most recent forecast should be regarded as a replacement for the first. Nonetheless, participants appeared to be using the inconsistency between forecasts to gage uncertainty, estimating wider ranges of outcomes than when forecasts were consistent. Moreover, like Experiment 2, participants appeared to use inconsistency to inform their decision strategy, closing schools more often, especially when forecasts were below the decision threshold.

General Discussion

These three experiments, the first specifically designed to examine the effects of sequential forecast inconsistency on trust, suggest that policies in favor of maintaining forecast consistency may be unwarranted. 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 far 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.

In order to ensure that our effects were due to the primary independent variables, inaccuracy and inconsistency, we attempted to control several 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 both forecasts to the outcome. Most but not all of these variables could be controlled in any given study, so we traded them across studies. Nonetheless, the basic results held in all experiments demonstrating their robustness and verifying that the effects reported here are due to inconsistency and inaccuracy per se, rather than to extraneous variables.

Across all three experiments, both inconsistency and inaccuracy were found to have a negative effect on trust. However, in each experiment the impact of inaccuracy was greater than that of inconsistency. In Experiments 1 and 3, the interaction of accuracy and consistency suggested that consistency matters mainly when forecasts are accurate. In addition, in Experiments 1 and 2, for the trials representing the options that meteorologists might most often face, inconsistent forecasts that were accurate were given higher trust ratings than consistent forecasts that were inaccurate. These results suggest that the gain in trust from maintaining

consistency may well be lost if the forecast turns out to be inaccurate. Importantly, and somewhat surprisingly, all three experiments provided evidence that consistency has an even smaller impact on pre-outcome trust, when participants were in process of making their decisions and accuracy is unknown. In other words, forecast inconsistency matters little even in the absence of accuracy information.

Additionally, there was no evidence that the effect of consistency builds over trials. We expected an increase in trust when forecast providers delivered consistent forecasts over a block of trials. However no such effects, in either direction, were detected. If the effect of consistency were building over the course of a block, the average trust in consistent trials would increase and the average trust in inconsistent trials would decrease over the course of a block. This was not the case overall. This may suggest that, to the degree that participants' trust was affected by forecast consistency, they regarded it as a characteristic of the forecast rather than the forecast provider.

It is also clear that inaccuracy significantly decreases trust. This effect was found across all three experiments, regardless of the stimuli that varied across them. Indeed the prominence of accuracy 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.

In addition to the impact of inconsistency and inaccuracy, we were interested in how people integrate information from differing forecasts. In all experiments, the weighting of the second forecast was at least 7 times greater than the earlier forecast regardless of whether or not

a forecast reminder was included. This may have been due to extra-experimental experience with real weather forecasts about which participants have many intuitions, often valid (Savelli & Joslyn, 2012). Perhaps people have noticed that more recent forecasts tend to be more accurate. It could also have been because our forecast data were realistic in the sense that forecasts grew more accurate as lead times decreased. That is, the first forecasts were accurate only 25% of the time while second forecasts were accurate 50% of the time. Participants might have learned (explicitly or implicitly) to discard 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 mode of presentation (simultaneous vs. sequential) or context (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 people understand that the more recent forecast is likely to be more accurate (which it generally is) so the difference across forecasts may matter less to them.

In addition, failing to update the forecast deprives users of potentially critical decision relevant information. In two of the three studies, people made more cautious decisions when forecasts were inconsistent, especially when the forecast predicted low snow totals, below the decision threshold. This suggests that inconsistency may be interpreted as an indication of uncertainty. Indeed, all 3 Experiments demonstrated that people expect a larger range of outcomes when forecasts were inconsistent relative to when they are consistent. This may in turn, encourage decision-makers to be more cautious in this scenario in which the cost of a miss far outweighs the cost of a false alarm. Therefore, not only is inconsistency less deleterious to

trust than inaccuracy but it may also provide the user with important information about potential uncertainty in the weather situation that may inform decisions that are better tailored to the users own risk tolerance.

This is not to say 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; etc.). However, when it comes to consistency in forecast content, since prioritizing consistency likely means deprioritizing accuracy as the work reported here demonstrates, the costs far outweigh the benefits.

These results have implications for a broad range of domains in which the accuracy of information is enhanced as lead times decrease. They suggest that although inconsistency in information can have a negative effect on trust, the 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 evidenced by the fact that our participants' expectations were influenced far more by the second than the first forecast. Thus, updating forecasts, even at a sacrifice to consistency, can preserve trust in the information source.

References

Budescu, Rantilla, Yu, & Karelitz. (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.

Eiser, J. R., & White, M. P. (2005, December). A psychological approach to understanding how trust is built and lost in the context of risk. In *CARR conference 'Taking Stock of Trust', London, London School of Economics* (Vol. 12).

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.

Gupta, N., Bisantz, A. M., & Singh, T. (2001, October). Investigation of factors affecting driver performance using adverse condition warning systems. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 45(23), 1699-1703. Los Angeles, CA: SAGE Publications.

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.

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.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-292.

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.

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.

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

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

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

	Experiment 1							
	Accurate			Inaccurate				
	Day 1	Day 2	Outcome	Day 1	Day 2	Outcome	Experiment 2	Experiment 3
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	5	7	3	7
	5	7	7	5	4	6	2	6
	6	4	4	6	7	5	9	5
	7	5	5	7	6	4	8	4

Note. Red values highlight differences in forecast values across experiments. All other values were the same across all experiments.

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	3.35	2.93	3.21	3.07	3.45	3.25	2.96	2.90
	Std error	.04	.05	.04	.05	.05	.05	.05	.05
	Std dev	.79	.904	.82	.88	.86	.91	1.01	1.01
2	Mean	3.42	3.03	3.32	3.12	3.54	3.29	3.10	2.96
	Std error	.06	.07	.07	.07	.07	.07	.08	.08
	Std dev	.78	.93	.85	.88	.86	.90	1.05	1.01
3	Mean	3.45	3.00	3.41	3.05	3.70	3.21	3.11	2.89
	Std error	.06	.07	.07	.07	.07	.07	.09	.07
	Std dev	.80	.90	.87	.83	.91	.88	1.08	.91

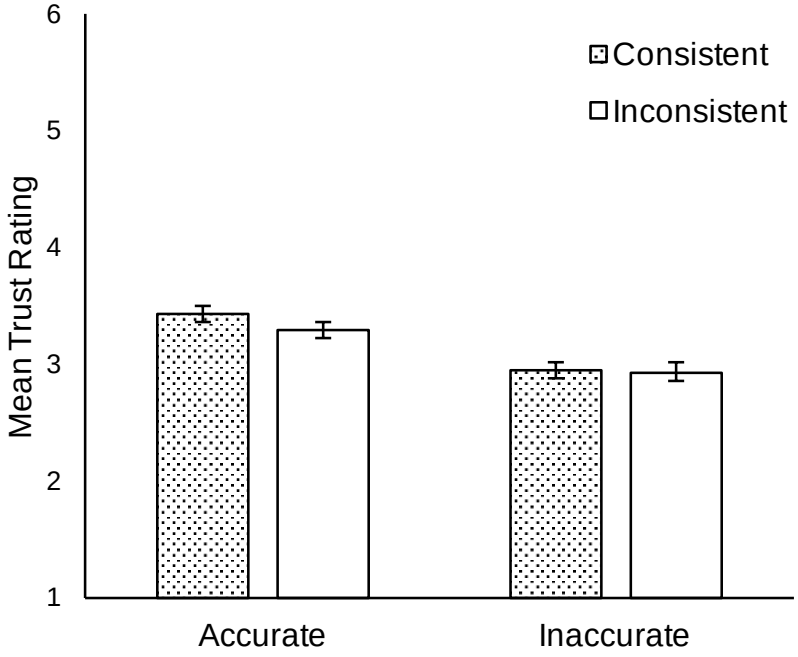

Figure 1. Experiment 1 Trust Ratings by Accuracy and Consistency.

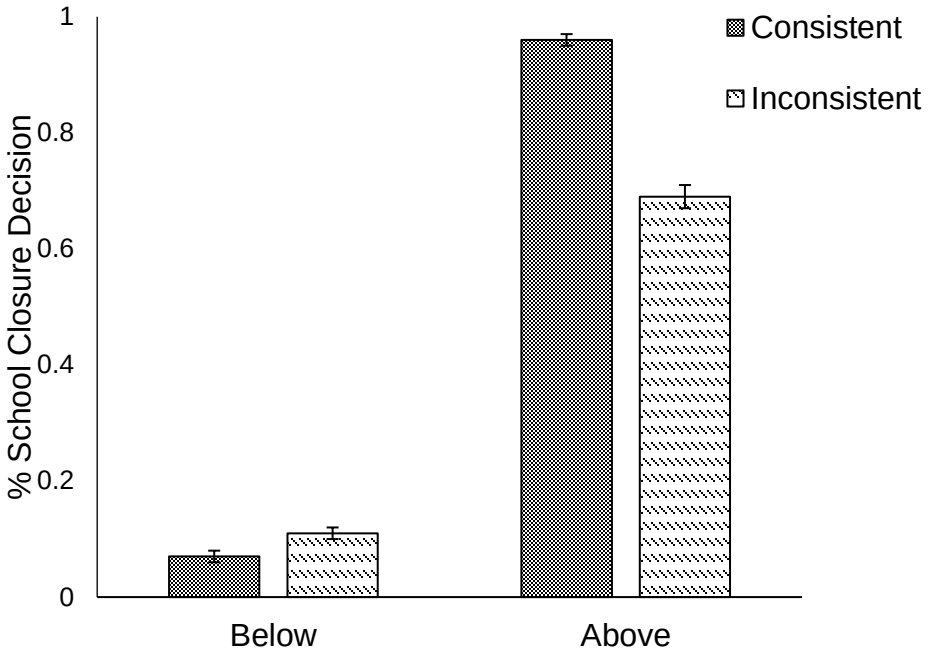

Figure 2. Experiment 1 Percent Closed by Threshold Orientation and Consistency

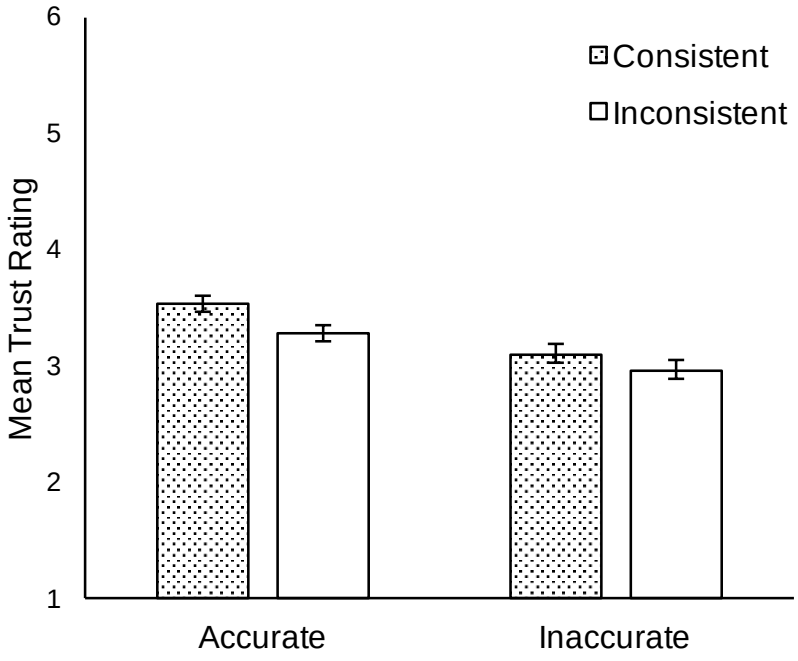

Figure 3. Experiment 2 Trust Ratings by Accuracy and Consistency.

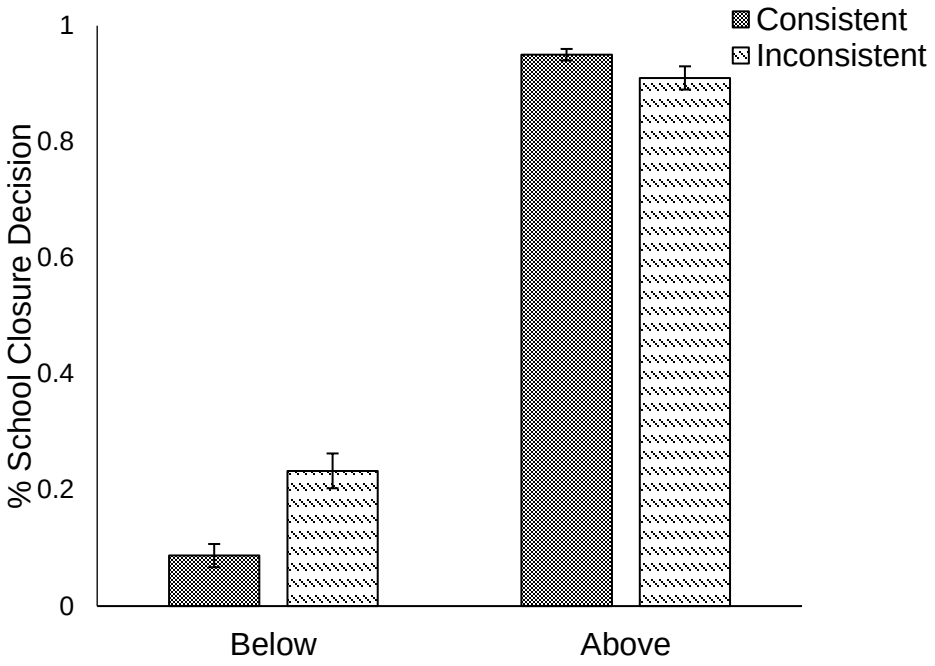

Figure 4. Experiment 2 Percent Closed by Threshold Orientation and Consistency.

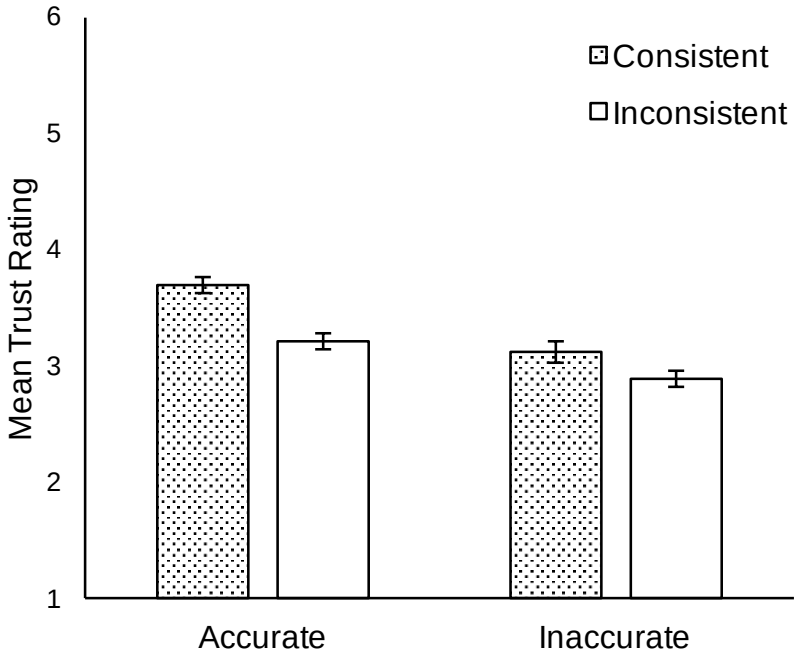

Figure 5. Experiment 3 Trust Ratings by Accuracy and Consistency

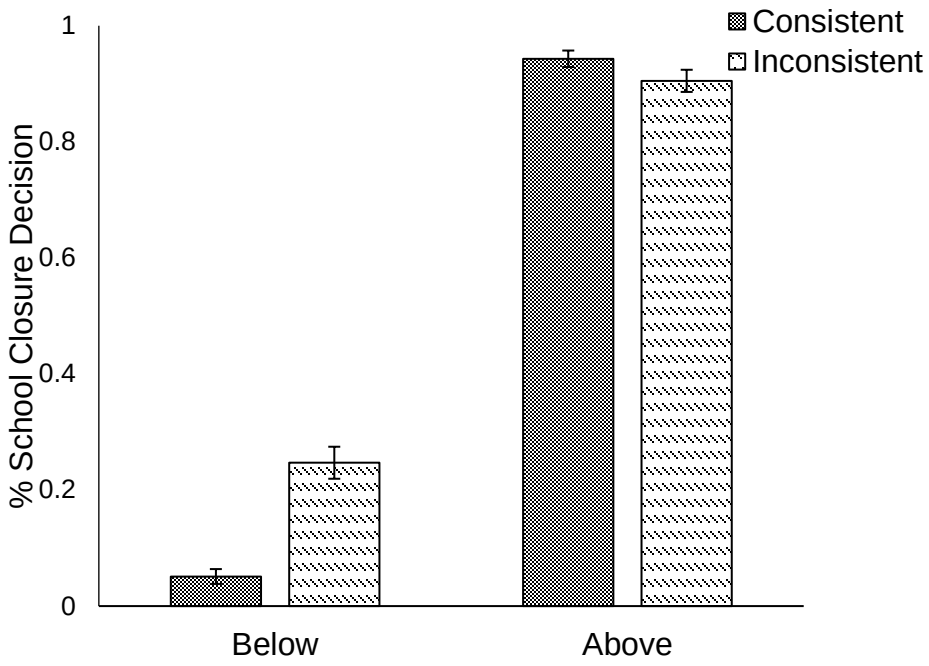

Figure 6. Experiment 3 Percent Closed by Threshold Orientation and Consistency.