Can Causal Sense-Making Benefit Foresight, Rather than Biasing Hindsight?

Edward Munnich (emunnich@usfca.edu)
Jennifer Milazzo (jlmilazzo@usfca.edu)
Jade Stannard (jostannard@usfca.edu)
Katherine Rainford (kprainford@usfca.edu)
Department of Psychology, 2130 Fulton St.
San Francisco, CA 94117 USA

Abstract

Upon reading headlines like "Traffic Fatalities Increased/Decreased Last Year," people often overestimate how well they would have anticipated changes. This hindsight bias has been linked to causal sensemaking that minimizes one's feeling of surprise after learning an outcome. In this paper, we consider whether the sensemaking process, which contributes to bias in hindsight, could be recruited to our benefit in foresight. We found that 1. Foresight participants—who estimated fatality statistics and listed causal factors before learning true statistics—were more surprised than Hindsight participants—who listed causal factors only after learning true statistics. 2. To the extent that Foresight participants were successful in listing causal factors in the *opposite* of their expected direction, they showed improvement in a second set of estimates they made prior to learning the true statistics; however, this improvement did not correspond to decreased surprise when they learned true statistics. We discuss implications for contrast vs. uncertainty theories of surprise, and for the possibility of useful belief revision triggered by unexpected statistics and consideration of alternative causation.

Keywords: causal reasoning, consider the opposite, explanation, foresight, hindsight bias, sensemaking, surprise

In 2005, there were 145 traffic fatalities per million Americans. Before reading further, please estimate how many traffic fatalities there were five years later in 2010. Next, think to yourself what factors caused traffic fatalities to increase or decrease from 2005 to 2010. Base rate statistics regarding trends in public safety, the economy, and the environment are readily available in the media or by searching online, and estimating base rates before receiving surprising feedback has been shown to affect personal and public policy preferences (e.g., Munnich, Ranney, Nelson, Garcia de Osuna, & Brazil, 2003; Ranney, Cheng, Nelson, & Garcia de Osuna, 2001). In particular, people find base rate statistics relevant when they have a compelling causal explanation for the causal mechanism behind the numbers (Tversky & Kahneman, 1982). Furthermore, base rates inform our actions to the extent that a causal explanation implicates a particular action—for example, Hagmayer and Sloman (2009) found people to be more likely to plan an action (e.g., recommend that a friend do more chores) when a statistic was presented along with a direct cause (e.g., people who do chores are healthier because they get exercise doing chores) as opposed to a common cause (e.g., people who do chores more are more conscientious, which also leads them to take better care of their health).

Now, please recall your estimate of 2010 traffic fatalities. In fact, auto fatalities declined from 145 per million US residents in 2005 to 106 per million in 2010. To the extent that the 2010 statistic is surprising to you, the surprise may prompt you to revise your beliefs about what contributes to traffic fatalities, and what factors can mitigate fatalities. On the other hand, had we presented the actual statistic at the beginning of the paper, many readers would have shown hindsight bias—the tendency to overestimate how well one would have anticipated an outcome before learning it (e,g., Fischhoff, 1975). A major contributing factor to hindsight bias is causal sensemaking (Ash, 2009; Pezzo, 2003; Roese & Olson, 1996; Schkade & Kilbourne, 1991): When we encounter an outcome that we would not have expected, if we are able to find a coherent explanation of the outcome, we often feel that we would have expected it all along.

This paper examines our ability to engage in a sensemaking process *before* learning an unexpected outcome—ideally at a point when our actions could still make a difference. In other words, we ask whether people can turn a process that amounts to a bias in hindsight into a benefit in foresight? To illustrate, if it starts to rain on the walk to work, one might remember a forecast a few days earlier that a storm was moving into the area—making sense of the drops falling on one's heads—and falsely reason that one expected it to rain (hindsight bias). But how could one have invoked the forecast at a point when one could have brought an umbrella and avoided getting wet? One way to accomplish this is by considering the opposite of an expected outcome in foresight, which has been shown to reduce bias (Slovic & Fischhoff, 1977; Lord, Lepper, & Preston, 1984; cf. Ranney, Rinne, Yarnell, Munnich, Miratrix, & Schank, 2008). In fact, one need not even consider the exact opposite outcome: Hirt and Markman (1995) found that just considering causes for a salient alternative to the expected outcome is sufficient to trigger debiasing. Obviously, sensemaking can only be as good as the knowledge one has, but, in principle, one should have access to all of the causal explanations before learning an outcome, that one would be able to think of immediately after learning an outcome. If one could think of these explanations a bit earlier, they could presumably benefit foresight, rather than biasing hindsight.

Were sensemaking solely focused on weighing the relative contributions of causal factors, consideration of the opposite outcome should move participants' second estimates (if at all) in the opposite direction of their original estimates. However Sanna, Schwarz, and colleagues (e.g., Sanna, Schwarz, & Small, 2002) found that metacognition also plays an important role: When participants did easy causal explanation task (e.g., thinking of *two* reasons why the British would have won a particular war), they rated outcomes to be more likely; however when the task was more difficult (as confirmed by subjective difficulty ratings; e.g., thinking of *ten* reasons why the British won), they actually rated the outcome to be *less* likely. With this in mind, our **Difficulty-Modulated Expectation Hypothesis** states that, to the extent that it is easy for one to think of explanations for an alternative outcome prior to learning the actual outcome, one's estimate will move in the opposite direction of one's initial estimate. Conversely, to the extent that one finds it difficult to think of explanations, one's estimate will become more polarized in the direction of the original estimate.

Another aspect to consider regarding moving sensemaking into foresight is how surprised one would be by the actual outcome upon learning it. Teigen and Keren (2003) cited evidence that people were not necessarily equally surprised by outcomes that they believed to be equally probable, and proposed the *contrast* hypothesis of surprise, in which one is surprised to the extent that an outcome contrasts with one's expectations. However, Maguire, Maguire, and Keane (2011) found that people were more surprised by an unexpected outcome when they had generated their own explanation, than when they were given the same explanation for the outcome. Notably, the unexpected event was equally contrastive with expectations in both cases, and the explanations that participants received were those most commonly generated by other participants—what apparently reduced surprise was the degree of *uncertainty* associated with generating an explanation oneself, as opposed to receiving an explanation. This leads to our **Reduced Surprise Hypothesis:** To the extent that one improves upon an estimate by considering an alternative outcome, one should be less surprised upon learning that the alternative outcome actually occurred. This provides a second kind of test of a contrast hypothesis against an uncertainty hypothesis of surprise—whereas Maguire et al. manipulated the certainty of the explanation, the present experiment manipulates the certainty of the outcome itself. If one's surprise is reduced by considering an alternative outcome, it would be consistent with a contrast hypothesis (i.e., the contrast was reduced, leading to a reduction in surprise), but if one's surprise is unaffected by an improvement in one's estimate, it would suggest that the uncertainty associated with generating an outcome oneself was responsible for the feeling of surprise.

Method

Participants and Design

98 participants were recruited through Amazon Mechanical Turk. The experiment used a 2x2 between-subjects design. The first independent variable was *prior estimation*: Hindsight participants received the true 2010 traffic fatality statistic at the start of the experiment, whereas the Foresight Group estimated 2010 statistic before receiving the true statistic. The second independent variable was *considering the opposite*: Consider Opposite (CO) participants were asked to consider factors that would cause fatality statistic to move in the opposite of the direction they expected (in Foresight, the opposite direction of their estimates, in Hindsight, the opposite direction of the true number) vs. those who did not consider the opposite; Non-Consider Opposite (NCO) participants skipped this step. Dependent variables were number of factors listed at each stage, estimates and re-estimates of 2010 traffic fatalities, surprise upon learning the actual 2010 fatality statistic, estimates of 2015 traffic fatalities, and estimates of how low traffic fatalities could go if all actions the participant suggested were implemented. With the exception of the number of factors that participants listed, all of these variables are ordinal (i.e., equal intervals cannot be assumed), so we used non-parametric statistical tests.

Materials and Procedure

Estimates. All Foresight participants received the actual statistic for 2005 traffic fatalities (i.e., 145 out of every million Americans were killed in car accidents in 2005) as a reference point, and they estimated how many Americans were killed in car accidents in 2010. Next, participants were prompted to describe up to five factors that they believed to have caused the change in U.S. traffic fatalities between 2005 and 2010. At all stages of the experiment that elicited causal factors, participants had to describe at least one factor to continue the experiment; after that, they were prompted to describe additional factors until they either reached five factors or selected "I cannot think of another factor".

Foresight + Consider Opposite. A subset of Foresight participants were then asked to imagine that the traffic fatality statistic actually moved in the *opposite direction* of what they predicted and to describe up to five factors that would have contributed to this change.

Re-Estimation. All Foresight participants (both CO & NCO) estimated the 2010 statistic a second time.

Incorporation of 2010 Statistic. Hindsight participants entered the experiment at this stage. All participants received statistics for *both* 2005 and 2010 (i.e., 106 out of every million Americans were killed in car accidents in 2010) to incorporate, and were asked to choose the statement that indicated how surprised they were with the change in the statistic between 2005 and 2010 ("not at all surprised"…"extremely surprised"). Subsequently, all four groups provided up to five factors they believed to have caused this decrease in U.S. traffic fatalities between 2005 and 2010.

Hindsight + Consider Opposite A subset of Hindsight participants were then asked to imagine that traffic fatalities had actually *increased* between 2005 and 2010 (i.e., in contrast to the statistics they had just received showing a decrease), and to describe up to five factors that could have caused such a change.

2015 Estimate All four groups provided estimates of what the traffic fatality rate *would be* in 2015, and described up to five factors they believed would affect traffic fatality rates between 2010 and 2015.

Actions To Reduce 2015 Fatalities Finally, all four groups listed up to five actions that could be taken to reduce traffic fatalities by 2015 (the experiment was carried out in the fall of 2011), and made a final estimate of how low the fatality rate could be if all actions were taken.

Coding Scheme

Each causal factor listed at each stage (Initial Estimation, Consider Opposite, Incorporation), was coded according to 17 possible categories of factors that plausibly cause either an increase or decrease in car accident fatalities (e.g., cell phone use, safer cars). We used a separate, binary code to indicate the direction of the effect—whether it was believed to have led to an increase or decrease in fatalities. This method allowed us to follow and closely analyze common causal threads through the experiment. For example, if a participant initially indicated that increased cell phone use contributed to an *increase* in fatalities, then later indicated that enforcement of laws against cell phone use contributed to a *decrease* in fatalities, our main code preserved the core idea that cell phones are a cause of fatalities. Two coders—who did not communicate about the experiment, and one of whom was unfamiliar with our hypotheses—coded items in opposite random orders. Interrater agreement was 85%, and, when there was disagreement between the coders, a third coder decided between the codes assigned by the first two coders. When participants listed two or more distinct factors within a single response field (e.g., "talking on cell phones and texting," where "and" clearly denotes a conjunction of two separate responses), we assigned a separate code for each distinct response.

Results

Estimates of 2010 Fatalities

At the outset, Foresight participants in what would become the CO (n=30) and NCO (n=23) groups listed similar numbers of unique factors (M_{CO} = 2.87, M_{NCO} = 2.74, t(51) = .36, n.s.; note: participants occasionally listed the same factor twice at one stage of the experiment—if so, we counted it as one unique factor). Both groups also provided similar types of causal factors—the top three factors affecting traffic fatalities across groups were driving under the influence of alcohol (CO:47%(14), NCO:43%(10)), cell phone use while driving (CO:40%(12), NCO:35%(8)), texting while driving (CO:40%(12), NCO:35%(8); since participants typically gave more than one response, percentages add to greater than 100%; no other factor was cited by more than 30% of participants).

The primary difference among Foresight participants was that 79% (n=42) of them estimated that fatalities increased between 2005 and 2010 ("Increasers": Median=217, Min=147, Max=800) and the remainder (n=11) estimated that fatalities decreased ("Decreasers": Median=120, Min=10, Max=142). This percentage replicated the roughly 80% of participants in pilot tests during the same time period, who believed that the traffic fatality rate was increasing. The consistently high percentage of people who shared this misconception made traffic fatality statistics a desirable focus for an investigation of surprise. Increasers listed numerically more factors than did Decreasers, but this did not reach significance (M_{Incr} = 2.90, M_{Decr} = 2.45, t(51) = 1.07, n.s.). However, there were differences in the types of factors listed: The factors listed most commonly by Increasers were driving under the influence of alcohol (50%(21)), cell phone use (48%(20)), and texting (45%(19)); by contrast, the factors most commonly listed by Decreasers were safety features of cars (82%(9)) and (un)safe driving (45%(5); no other factor was cited by more than 30% of participants).

Consider the Opposite (CO) vs. Not (NCO)

A subset of Foresight participants (as it turned out, seven were Decreasers and 23 were Increasers) were then asked to consider the opposite. CO-Decreasers were asked what could have caused an *in*crease in fatalities, and CO-Increasers were asked what could have caused an *de*crease in fatalities. Each group listed roughly the same number of unique factors (M_{Decr}=2.43, M_{Incr}=2.57). The patterns of additive factors (i.e., not mentioned earlier; M_{Decr}=1.57, M_{Incr}=1.13) and subtractive factors (i.e., reversing the direction of factors that were given earlier when they estimated—e.g., if they thought cars had gotten safer, now they would say cars had gotten less safe; M_{Decr}=0.86, M_{Incr}=1.43), differed somewhat between Increasers and Decreasers, but neither of the main effects, nor Group x Response Type interaction were significant. When they considered the opposite—an increase in fatalities—Decreasers continued to cite (un)safe driving as a major factor (57%(4)), but now mentioned cell phone use (43%(3))—which none of the Decreasers mentioned when they initially estimated, but which was a major factor for Increasers at that stage—and number of drivers on the road (43%(3)) moved into their top three factors. When considering the opposite, Increasers continued to cite alcohol (43%(10))—now, as a decrease in driving under the influence—but also mentioned (un)safe driving (also 43%(10)), which none of them had mentioned when they estimated but which had been a major factor among Decreasers. In other words, both groups showed a mix of reversing their earlier explanations (subtractive factors) and invoking new explanations (additive factors) when they considered the opposite. To the extent that they cited additive factors, they moved towards the patterns cited by the other group at the estimation stage.

Re-estimation of 2010 Fatalities

After CO participants considered reasons for changes in the opposite direction, *all* Foresight participants re-estimated 2010 fatalities. Among Decreasers, all four NCO participants and four out of seven CO participants gave identical responses when they re-estimated. Of the remaining COs, two made small adjustments (<8 fatalities per million), and only one showed a notable change from estimate to re-estimate (10→125 fatalities per million), which is consistent with the hypothesis that considering the opposite would lead to more accurate estimates.

Among Increasers, CO participants (n=23) were both more likely to change their re-estimates of 2010 fatalities ($X^2(1) = 4.71$, p = .03), and the changes in their re-estimates were larger in magnitude ($\text{Median}_{CO} = 15$), than those of NCO participants (n=19; $\text{Median}_{NCO} = 0$; Mann-Whitney $U = 139$, $p = .03$). Moreover, among CO participants who initially estimated an increase in fatalities, the more factors they listed for a decrease in fatalities when considering the opposite, the more their re-estimates decreased (i.e., improved; $r_{Spearman} = -.40$, $p_{\text{one-tailed}} = .028$; Figure 1). That said, the changes in re-estimates were not necessarily improvements: Although 10 participants' re-estimates decreased (improved) and eight were unchanged, five participants' re-estimates *increased* (moved further from the actual statistic). Together with the correlation between number of factors and change in re-estimate, we see that those who provided the fewest factors when considering the opposite tended to move away from the actual statistic, as predicted by the difficulty-modulated aspect of the Difficulty-Modulated Expectation Hypothesis.

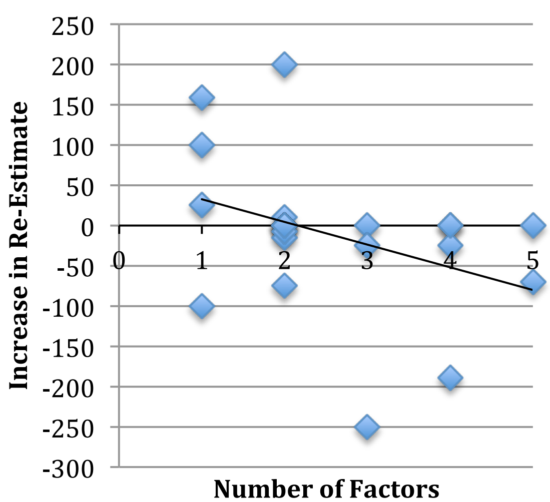

Figure 1. Correlation between number of factors listed when considering a decrease and increase in re-estimate, among those who initially estimated that fatalities would increase. Negative values denote re-estimates that were smaller than initial estimates (i.e., *closer* to actual value).

Incorporation of 2010 stats

Upon learning the actual statistic, Foresight participants were reliably more surprised (Median=3, "very surprised") than Hindsight participants (Median=1, "slightly surprised", Mann-Whitney $U = 633$, $n_{\text{Foresight}}=53$, $n_{\text{Hindsight}}=45$, $p < .001$; Figure 2). Apart from the difference in level of surprise, the groups differed in what best predicted their surprise: Among Hindsight participants, the more factors one listed to explain the actual fatality statistic, the less surprised one was ($r_{\text{Spearman}}=-.51$, $p<.001$); by contrast, there was no correlation between factors listed and surprise among Foresight participants ($r_{\text{Spearman}}=.00$, n.s.). However, Foresight participants' surprise was well predicted by both their initial estimates ($r_{\text{Spearman}}=.68$) and re-estimates ($r_{\text{Spearman}}=.69$, both $ps<.001$). Notably, among Increasers who considered the opposite—that is, those whose initial estimates were in the wrong direction, but who then had the chance to improve theie estimates by thinking of alternative factors—neither the number of CO factors listed ($r_{Spearman} = -26$, $p = .11$, one-tailed) nor their improvement from estimate to re-estimate ($r_{Spearman} =-.02$, n.s.), reliably predicted their surprise.

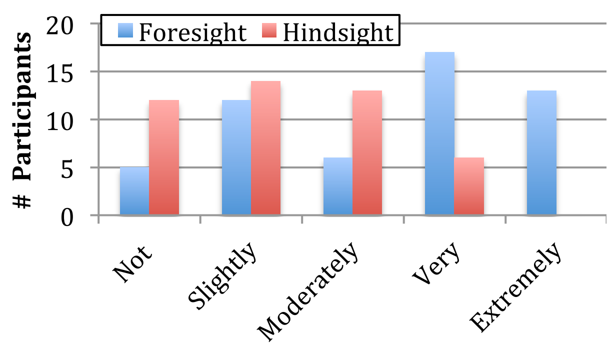

Figure 2: Frequency of surprise responses of Foresight vs. Hindsight Groups—ranging from "not at all surprised" to "extremely surprised."

Expectations for 2015

Foresight participants' estimates for 2015 fatalities were slightly, but reliably lower than Hindsight participants' (Mann-Whitney U=920, $n_{\text{Foresight}}=53$, $n_{\text{Hindsight}}=45$, p=.049). Looking more closely, there was no difference between Hindsight and Foresight participants who considered the opposite ($\text{Median}_{\text{Foresight+CO}} = 98$, n=30; $\text{Median}_{\text{Hindsight+CO}} = 99$, n=21; $U = 295$, n.s.). However, of those who did *not* consider the opposite, Foresight participants provided reliably lower estimates ($\text{Median}_{\text{Foresight+NCO}} = 95$, n=23) for 2015 traffic fatalities than Hindsight participants ($\text{Median}_{\text{Hindsight+NCO}} = 100$, n=24; $U = 168$, $p = .02$). Surprise regarding 2010 statistics was not a predictor of 2015 estimates for either the Hindsight ($r_{\text{Spearman}} = .05$) or Foresight group ($r_{\text{Spearman}} = -.04$), nor were 2010 estimates or re-estimates by the Foresight group reliable predictors of their 2015 estimates (Estimate: r_{Spearman} = -.14; Re-Estimate: r_{Spearman} = -.22, all n.s).

Actions to Reduce Traffic Fatalities in 2015

Overall, Foresight participants provided numerically lower estimates than Hindsight participants for 2015 fatalities if all actions they specified were taken, but the difference did not reach significance (Mann-Whitney U=990, $n_{\text{Foresight}}$=53, $n_{\text{Hindsight}}$=44, p=.20). Looking more closely, there were no reliable differences between participants who considered the opposite ($\text{Median}_{\text{Foresight+CO}}$ = 85; $\text{Median}_{\text{Hindsight+CO}}$ = 88; Mann-Whitney U =268, n.s.), and, among those who did not consider the opposite, there was a directional trend that mirrored the differences between Hindsight and Foresight groups' 2015 estimates, but this failed to reach significance ($\text{Median}_{\text{Foresight+NCO}}$ = 80; $\text{Median}_{\text{Hindsight+NCO}}$ = 85; Mann-Whitney U=216, p =.28). Surprise response to 2010 statistics was not a predictor of how low participants believed 2015 fatality rates could go for either the Hindsight (r_{Spearman} = .15) or Foresight group (r_{Spearman} = -.21), but, interestingly, there was a reliable trend such that the higher Foresight participants' 2010 estimates and re-estimates, the lower they thought 2015 fatalities could go if all actions they specified were taken (Estimate: r_{Spearman} = -.41, p=.002; Re-Estimate: r_{Spearman} = -.39, p=.004).

Discussion

The present findings are consistent with the Difficulty-Modulated Expectation Hypothesis—participants who considered the opposite showed improvement between their estimates and re-estimates of 2010 traffic fatalities to the extent, in proportion to the number of reasons they could think of for traffic fatalities to move in the opposite of their expected direction. In fact, those who thought of the fewest reasons apparently perceived the difficulty of thinking of opposite factors as making the opposite less likely, and moved further from the actual traffic fatality statistic when they considered the opposite (echoing Sanna et al., 2002).

At the same time, the results were not consistent with the Reduced Surprise Hypothesis: Whereas those who listed the most factors for the opposite of their expected direction for 2010 fatalities showed the most improvement in their re-estimates, they showed no corresponding reduction in the surprise they felt learning the actual statistics. This result is inconsistent with a characterization of surprise as the product of contrast (Teigen & Keren, 2003), but it provides converging evidence for Maguire et al.'s (2011) characterization of surprise as the result of uncertainty. Whereas Maguire et al. found that surprise corresponded to the uncertainty of the explanation for an unexpected outcome, the present study extends the breadth of evidence by supporting a parallel phenomenon regarding the level of certainty one attaches to the outcome itself. In short, even when our beliefs ultimately prove to be accurate, we do not view them with the same level of certainty as outcomes that are presented to us as fact. One caveat is that those who were able to list more factors did show a trend in the direction of lower surprise, and perhaps would show a significant relationship with greater power. But even if that were the case, the relationship between number of factors and surprise would be much weaker than that between number of factors and improvement in re-estimates. Moreover, the lack of any correlation between improvement and surprise suggests that surprise is, in any case, not merely a product of the contrast between what one expects and what one learns to be true.

Although improvement in estimates did not reduce surprise, we did find that merely being in the Foresight group led to greater surprise. This is also consistent with an uncertainty explanation of surprise—thinking about a statistic before learning it brings a level of uncertainty that one does not feel in hindsight, and this corresponds to a level of surprise. Interestingly, although surprise was strongly related to the number of causal factors Hindsight participants thought of (a replication of hindsight bias effects), there was no relationship between the number of factors offered by Foresight participants when they learned the actual 2010 fatality statistics and their surprise; rather the Foresight group's surprise was a function of how accurate their initial estimates, and their re-estimates were. Initial Foresight condition also played a role in estimates of 2015 fatalities—Foresight participants who did *not* consider the opposite provided reliably lower estimates than did Hindsight participants who did not consider the opposite, and a similar trend was present for estimates of how low 2015 fatalities could go if all actions a participant suggested were taken, but it did not reach significance. Interestingly, what did predict how low 2015 fatalities could go was how *high* one's initial estimate was. Mindful of the fact that these are post-hoc observations, we speculate that something interesting takes place when sensemaking regarding alternatives is not in the picture: The initial way that one approaches the question is important, and engaging in Foresight may lead one to more radical predictions for the future (of course, we do not know whether these predictions are correct yet). It appears that just thinking about base rate statistics in foresight—even if we cannot (e.g., CO participants whose surprise was no greater if they thought of fewer reasons), or do even try to think of alternative factors (e.g., NCO participants) can inoculate us from hindsight bias. That is, by estimating ahead of time, we are reminded of what we *did not know* prior to learning the statistic. This is consistent with the different patterns of allocation of health care spending seen in those who engaged in foresight about disease prevalence (Rinne, Ranney, & Lurie, 2006), and might contribute to the potential for belief revision with discrepant events (Clement & Steinberg, 2002; Ranney & Thagard, 1988) and surprising statistics (Munnich, Ranney, & Song, 2007).

Taken together, the results regarding our two main hypotheses portray both a similarity between the improvement in prediction that is possible when one moves sensemaking from being a biasing factor in hindsight to being a benefit in foresight. At the same time, there is a crucial difference: Sensemaking prior to learning an outcome does not reduce one's surprise in the way that sensemaking immediately after learning an outcome appears to—presumably because no amount of explanation in foresight can bring one the certainty that comes with knowing the actual statistic. One way of characterizing this is that cognitively, sensemaking before and after learning an outcome share strong similarities, but metacognitively, they are quite different. In any case, this is hopeful news if we aim to think of the causal explanations in foresight that would otherwise occur to us immediately after learning an an outcome. By considering alternative outcomes ahead of time, we can take useful action—like grabbing an umbrella in time to avoid getting wet, or supporting policies and adopting personal behaviors that could prevent traffic fatalities.

Acknowledgments

We are deeply grateful for the financial support of the USF Faculty Development Fund, and for the many helpful suggestions from Melissa Latham, Dorina Sandoval, Shaun Morris, Katheryn Conde, Caleb Banks, Alfred Si, Michael Stagnaro, Saera Khan, and Marisa Knight, Tania Lombrozo and the Concepts and Cognition Group at Berkeley, and Michael Ranney and the Reasoning Group at Berkeley—where many of these ideas were born and have been nurtured over the years.

References

Ash, I. K. (2009). Surprise, memory, and retrospective judgment making: Testing cognitive reconstruction theories of the hindsight bias effect. *Journal of Experimental Psychology: LMC, 35,* 916–933.

Clement, J. & Steinberg, M. (2002). Step-wise evolution of mental models of electric circuits: A "learning-aloud" case study. *Journal of the Learning Sciences 11*(4), 389-452.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, I,* 288-299.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, 138(1), 22-38.

Hirt, E., & Markman, K. (1995) Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69*, 1069-1086.

Lord, C., Lepper, M., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231-1243.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: LMC, 37*, 176-186.

Munnich, E.., Ranney, M., Nelson, J., Garcia de Osuna, J., and Brazil, N. (2003). Policy shift through Numerically-Driven Inferencing: An EPIC experiment about when base rates matter. Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society. (pp. 834-839). Mahwah, NJ: Erlbaum.

Munnich, E., Ranney, M., A. & Song, M. (2007). Surprise, surprise: The role of surprising numerical feedback in belief change. In D.S. McNamara & G. Trafton (Eds.) *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 503-508). Mahwah, NJ: Erlbaum.

Pezzo, M. (2003). Surprise, defence, or making sense: What removes hindsight bias? *Memory, 11*(4/5). 421-441.

Ranney, M., Cheng, F., Nelson, J., and Garcia de Osuna, J. (2001). Numerically driven inferencing: A new paradigm for examining judgments, decisions, and policies involving base rates. Paper presented at the Annual Meeting of the Society for Judgment & Decision Making.

Ranney, M., Rinne, L. F., Yarnall, L., Munnich, E., Miratirx, L., & Schank, P. (2008). Designing and assessing numeracy training for journalists: Toward improving quantitative reasoning among media consumers. In P. A. Kirschner, F. Prins, V. Jonker, & G. Kanselaar (Eds.), *International Perspectives in the Learning Sciences: Proceedings of the Eighth International Conference for the Learning Sciences* (pp. 2-246-2-253). Intl. Society of the Learning Sciences, Inc.

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. Proceedings of the Tenth Annual Conference of the Cognitive Science Society (pp. 426-432). Hillsdale, NJ: Erlbaum.

Rinne, L., Ranney, M., & Lurie, N. (2006). Estimation as a catalyst for numeracy: Micro-interventions that increase the use of numerical information in decision-making. In S. Barab, K. Hay, & D. Hickey (Eds.). *Proceedings of the Seventh International Conference of the Learning Sciences* (pp. 571-577). Mahwah, NJ: Erlbaum.

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology, 32,* 197–227.

Sanna, L. J., Schwarz, N., & Small, E. M. (2002). Accessibility experiences and the hindsight bias: I knew it all along versus it could never have happened. *Memory & Cognition, 30,* 1288–1296.

Schkade, D. A., & Kilbourne, L. M. (1991). Expectation-outcome consistency and hindsight bias. *Organizational Behavior and Human Decision Processes, 49,* 105–123.

Slovic, P., & Fischhoff, B. (1977) On the psychology of experimental surprises. *Journal of Experimental Psychology: HPP, 3*, 544-551.

Teigen, K. H., & Keren, G. (2003). Surprises: Low probabilities or high contrasts? *Cognition, 87*, 55–71.

Tversky, A. & Kahneman, D. (1982). Evidential impact of base rates. In: D. Kahneman, P. Slovic, and A. Tversky (Eds.) *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.