# A Rational Account of Anchor Effects in Hindsight Bias

**Samarie Wilson (samariew@princeton.edu)**
Mechanical and Aerospace Engineering, Princeton University
Princeton, NJ 08540 USA

**Somya Arora (somyaa@princeton.edu)**
Department of Computer Science, Princeton University
Princeton, NJ 08540 USA

**Qiong Zhang (qiongz@princeton.edu)**
Princeton Neuroscience Institute, Princeton University
Princeton, NJ 08540 USA

**Thomas L. Griffiths (tomg@princeton.edu)**
Departments of Psychology and Computer Science, Princeton University
Princeton, NJ 08540 USA

## Abstract

Hindsight bias is exhibited when knowledge of an outcome (i.e., an anchor) affects subsequent recollections of previous predictions (i.e., an estimate). Hindsight bias usually leads to estimates being remembered as closer to the anchor than they actually were. The exact amount of hindsight bias exhibited depends on the anchor value and the anchor plausibility, with experimental results showing that hindsight bias is elicited only when the anchor is perceived to be plausible. In this paper we present a Bayesian model that captures the relationship between hindsight bias and anchor plausibility. This model provides a rational account of hindsight bias by considering memory recall as a statistical problem, where the goal is to reconstruct the original estimate using the anchor as new evidence. Simulations show that the modeled trends align closely with previously published human data.

**Keywords:** hindsight bias; rational analysis; Bayesian inference; anchor plausibility; prior

## Introduction

Imagine you are at a soccer game with your friend. Before the game begins, you predict that your favorite team is going to win by 10 points. However, at the end of the game, they only win by 2 points. While discussing the game with your friend afterwards, you recall that you always knew it was going to be a close game. This difference between the original estimate, your prediction of a certain team winning by 10 points, and your recalled estimate, the game being a close one, is called hindsight bias – the recalled estimate is biased towards the outcome, known as an "anchor" (Synodinos, 1986; Tversky & Kahneman, 1974). Hindsight bias is also known as the "knew it all along" effect (Roese, 2012).

Previous studies show that hindsight bias does not always occur, but rather depends on the plausibility of the anchor (Wegener, Petty, Detweiler-Bedell & Jarvis, 2001; Chapman & Johnson, 1994). The plausibility can be manipulated in a continuous manner by varying the distance between anchor and the original estimate: an anchor is perceived as increasingly implausible as this distance increases (Hardt & Pohl, 2003).

Biases like the hindsight bias have typically been taken as evidence of human irrationality, but recent work has suggested that many apparently irrational biases can be made sense of within rational models (Lieder, Griffiths, Huys & Goodman, 2017; Feldman & Griffiths, 2007; Hemmer & Steyvers, 2009, Parpart, Jones and Love, 2018; Sher, McKenzie, 2006). Here we pursue a similar approach, providing a rational account of the hindsight bias.

We focus on explaining the plausibility effect. Specifically, we focus on modeling the experimental results in Hardt & Pohl (2003), as it provides continuous measures of anchor distances, which allows for examining a range of values on anchor plausibility. The exact relationship between hindsight bias and anchor distance is shown in Figure 1. As we can see in Figure 1, the extent an anchor influences memory depends on how much it is inferred to be plausible: when the anchor is considered to be plausible, the relationship between the anchor distance and hindsight bias is directly proportional. However, beyond a certain anchor distance, the anchor is considered to be more and more implausible, and the relationship is inversely proportional. Previous modeling work by Pohl, Eisenhauer & Hardt (2003) provides a mechanistic account of hindsight bias, but no current model accounts for the effects of anchor plausibility.

In this paper, we define and evaluate a Bayesian model of the relationship between hindsight bias and anchor plausibility. The paper is organized as follows. In the following section, we provide a brief overview of existing work regarding hindsight bias. Next, we develop the mathematical model for how the anchor and perceived plausibility update the prior. This sets us up to apply the formalization to simulate the relationship between hindsight bias and anchor distance as in Figure 1. Finally, we review the relationship between hindsight bias and anchor plausibility, restate our assumptions, and lay out future directions for this work.

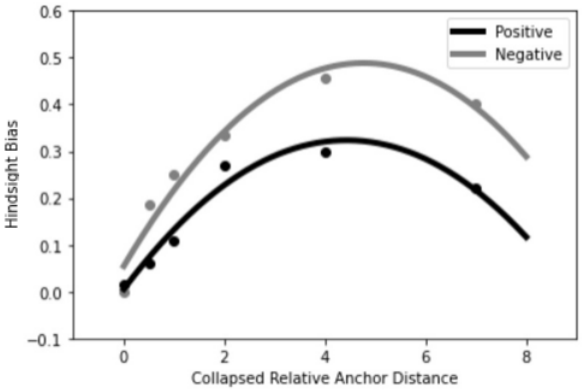


Figure 1: Hindsight bias and its dependence on anchor distance as reproduced from Figure 4 of Hardt & Pohl (2003). The magnitude of hindsight bias, shown on the vertical axis, is the amount that the revised estimate moves from the original estimate towards the provided anchor. Positive and negative labels indicate whether the anchor values were above or below the original estimate, respectively.

## Background: The Hindsight Bias in Hindsight

### Hindsight Bias Effect

Hindsight bias occurs when people feel that they "knew it all along" (Roese, 2012; Roese and Vohs, 2012). It is elicited in human memory by the presentation of an anchor. Estimates are generally recalled to be closer to the anchor than they had been in reality (Christensen-Szalankski & Beach, 1984), which has been postulated to take place either during recollection or during reconstruction of memory (Erdfelder, Brandt, & Broder, 2007). Meta-analysis across many studies has shown that effects of hindsight bias can be moderated by the type of anchor information presented and the subject's familiarity with the task. Observed inconsistencies in the effects of hindsight bias suggest that it may be a result of cognitive factors rather than motivational ones (Christensen-Szalanski & Willham, 1991). Overall, studies have underscored the robustness of hindsight bias (Christensen-Szalankski & Willham, 1991). Therefore it is worth exploring the manipulations that interrupt the effects of anchor presentation on recollection of the original estimate (Hawkins & Hastie, 1990).

### Effect of Anchor Plausibility

Previous studies have shown that anchor plausibility is a function of the anchor's agreement with the individual's prior knowledge of the relevant subject (Pohl, 1998; Strack & Mussweiler, 1997). Intuitively, invalidating anchor data undermines its functionality, thereby diminishing the hindsight bias (Hasher, Attig & Alba, 1981). Determining an anchor to be implausible is an example of such an invalidation. The two factors that describe anchor plausibility are known to be knowledge quantity and knowledge precision (Pohl, 1998).

Studies on the effect of anchor plausibility on hindsight bias have produced inconsistent findings. In one experiment, Strack and Mussweiler (1997) found no differences between the influences of plausible and implausible anchors. In another, Chapman and Johnson (1994) found the anchoring effect of extremely implausible anchors to be insignificant.

Hardt and Pohl (2003) revisited anchor plausibility, focusing on the idea that the subjective plausibility of anchors is the key variable. Using measures of the extent to which participants perceived continuously varying anchors as plausible, they were able to produce more nuanced results than previous studies. In their study, participants were given several questions (e.g., how old was Gandhi when he died?) and asked to give the minimum and maximum values that they considered to be plausible answers to each question, as well as an exact estimate of the answer. After a week, participants were brought back and presented with what they believed were other participants' estimates for each of the questions, before recalling their own estimates from the week prior. The influence of these presented estimates, (i.e., anchors), on the recollection depended upon how plausible the anchors were perceived to be by the participants.

### Existing Modeling

Hindsight bias has been modeled by Pohl, Eisenhauer and Hardt (2003) using the SARA model, a simplified and focused version of the associative memory model SAM (Search of Associative Memory; Raaijmakers & Shiffrin, 1980; Shiffrin & Raaijmakers, 1992). SARA is a process model designed to better understand each process of the hindsight bias phenomenon: estimate generation, anchor encoding, and estimate recollection. In this model, estimates are generated by a probabilistic sampling process that samples knowledge representations from long-term memory and compiles all these representations to create an estimate. Presented anchors are automatically encoded, and this encoding process strengthens the representations that are most like the presented anchor. This makes these representations more likely to be recalled in subsequent processes that use this knowledge to reconstruct the original estimate. The reconstruction of the original estimate is thus biased towards the anchor and representations similar to the anchor, producing hindsight bias. While SARA successfully models the hindsight bias processes, it does not capture the effect of anchor plausibility on hindsight bias (Pohl, Eisenhauer & Hardt, 2003).

## Hindsight as Bayesian Inference

Our model frames memory recall as a problem of statistical inference. The goal is to reconstruct the original estimate using the noisy memory representation of the estimate and the additional evidence from the anchor. It is rational to incorporate information from the anchor to the degree that it is plausible.

To express this model formally, we introduce variables $m$, $s$ and $a$, where $m$ is the original estimate, $s$ is the memory of

the original estimate, and $a$ is the presented anchor. The goal is to reconstruct $m$, based on the evidence provided by the noisy memory trace, $s$, and the anchor, $a$. $s$ is assumed to be centered around the true value of the original estimate, $m$:

$$p(s|m) \sim N(m, \sigma_s^2) \tag{1}$$

and $p(m)$ is assumed to be a uniform distribution. The anchor $a$ provides another source of information relevant to $m$, being an answer to the estimation problem. If $m$ is a good estimate of $a$, then the two should be related and $p(a|m)$ will focus on values close to $m$. Reconstructing $m$ is then simply a matter of Bayesian inference, with $p(m|a,s) \propto p(a|m)p(s|m)p(m)$.

However, there is a catch: this analysis assumes that $a$ is a plausible anchor. In practice, people infer whether an anchor is reasonable. To capture this we introduce a third variable, $z$, to represent the plausibility of $a$. The implausible case is denoted by $z = 0$, and the plausible case is denoted by $z = 1$. The two categories of anchors can be captured by defining $p(a|m,z)$ to be Gaussian distributions for each value of $z$. Intuitively, the larger the anchor distance, the lower the perceived plausibility. Therefore the plausible anchors are centered closer to $m$, whereas implausible anchors have a larger variance, making $p(a|m,z)$ look relatively uniform. This results in the distributions shown in Figure 2, which can be written as:

$$P(a|m,z) \sim \begin{cases} N(m, \sigma_0^2), & z = 0 \\ N(m, \sigma_1^2), & z = 1 \end{cases} \tag{2}$$

where $\sigma_0^2 >> \sigma_1^2$. This makes intuitive sense because in the plausible case ($z = 1$), the distribution will have larger probability values near the mean of the memory, whereas for the implausible case ($z = 0$), the distribution will be relatively flat. While $\sigma_0^2$ can be infinitely large, regardless what question is answered; $\sigma_1^2$ may be question dependent, related to the typical range of plausible answers to that question. We also assume that the prior probabilities assigned to the plausible and implausible categories, $p(z)$, are equal:

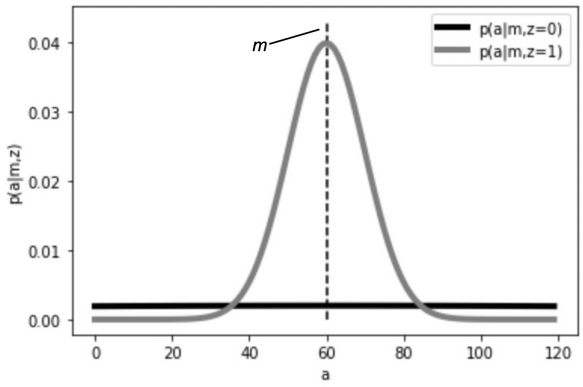


Figure 2: $p(a|m,z)$ for values of a with $\sigma_0 = 10$ and $\sigma_1 = 200$.

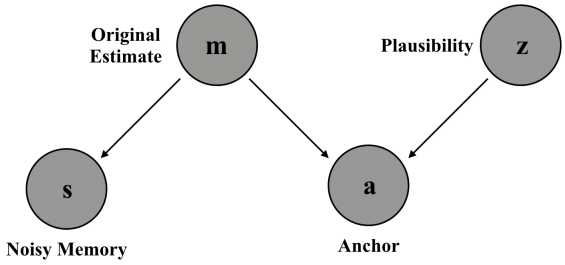


Figure 3: Generative Model for how the prior of original estimate, $m$, is updated using the anchor ($a$) and the plausibility ($z$).

$$p(z = 0) = p(z = 1) = \frac{1}{2}. \tag{3}$$

Given the aforementioned assumptions, we express the relationships between these variables using a generative model. Here the anchor, $a$, is produced either from a plausible source or implausible source, $z$, both centered at $m$. $s$ is a noisy representation of the original estimate $m$. The inference of plausibility and the incorporation of the anchor is determined by the distribution of original estimate, $m$, giving our generative model the structure shown in Figure 3.

Now that we have all our assumptions finalized for our model, we can define how we will quantify hindsight bias. The goal of our model is to determine how the anchor updates the memory, $p(m|a,s)$, and to use this to compute the hindsight bias. Once we have updated the memory, we can take hindsight bias, $\Delta h$, to be the difference between the posterior mean, $E(m|a,s)$, and the original estimate, $m$:

$$\Delta h = E(m|a,s) - m. \tag{4}$$

If the plausibility of the anchor is known, it is straightforward to infer $m$ by computing $p(m|s,z,a)$. However, in the case we don't know the plausibility of the anchor, we need to sum over the possible values of $z$ when computing $p(m|a,s)$:

$$p(m|a,s) = \sum_z p(m|s,z,a)p(z|a,s) \tag{5}$$

where $p(z|a,s)$ reflects the inferred plausibility for a given anchor and $p(m|s,z,a)$ is the posterior distribution over $m$ computed by assuming that the anchor is either plausible or implausible.

We can rewrite the posterior distribution by applying Bayes' rule:

$$p(m|s,z,a) \propto p(a|m,z)p(s|m)p(m) \tag{6}$$

Given that $p(a|m,z)$ and $p(s|m)$ are normally distributed as seen from Equations 1 and 2 and $p(m)$ being uniform, we can

write the posterior $p(m|s,z,a)$ as

$$p(m|s,z=0,a) \sim N\left(\frac{\sigma_s^2}{\sigma_s^2+\sigma_0^2}a+\frac{\sigma_0^2}{\sigma_s^2+\sigma_0^2}s, \left(\frac{1}{\sigma_s^2}+\frac{1}{\sigma_0^2}\right)^{-1}\right) \tag{7}$$

$$p(m|s,z=1,a) \sim N\left(\frac{\sigma_s^2}{\sigma_s^2+\sigma_1^2}a+\frac{\sigma_1^2}{\sigma_s^2+\sigma_1^2}s, \left(\frac{1}{\sigma_s^2}+\frac{1}{\sigma_1^2}\right)^{-1}\right). \tag{8}$$

Finally, the two components of the posterior are weighted by $p(z|a,s)$ in Equation 5. We can apply Bayes' rule to get:

$$p(z|a,s) \propto p(a|z,s)p(z) \tag{9}$$

We also have

$$p(a|z=0,s) \sim N(s, \sigma_s^2+\sigma_0^2) \tag{10}$$

and

$$p(a|z=1,s) \sim N(s, \sigma_s^2+\sigma_1^2) \tag{11}$$

which are derived by incorporating the additive noise from Equation 2 into the normal distribution in Equation 1.

Given a value of an anchor, we can calculate the exact values of $p(a|z=0,s)$ and $p(a|z=1,s)$ using Equations 10 and 11. Combining them with Equation 9 and the property that $p(z=0|a,s)$ and $p(z=1|a,s)$ sum to 1, we can find the exact values of $p(z=0|a,s)$ and $p(z=1|a,s)$. By substituting them into Equation 5, together with Equation 7 and 8, we obtain the updated memory, $p(m|a,s)$. Now that we understand how the anchor updates the memory, $p(m|a,s)$, we can compute the hindsight bias using Equation 4.

## Simulations

Having derived a Bayesian model of the effects of plausibility on hindsight bias, we now show that this model can reproduce the relationship between anchor distance and the magnitude of this bias shown in Figure 1. A foundation of the model is the relationship between the plausible and implausible distributions ($\sigma_0$ and $\sigma_1$). It is important that $\sigma_0 >> \sigma_1$. Plausible anchors are centered closely around the original estimate, $m$, and implausible anchors are further away from it.

In order to simulate hindsight bias, which can be written as $E(m|a,s)-m$, where $p(m|a,s)=\sum_z p(m|s,z,a)p(z|a,s)$, we must first find $p(m|s,z,a)$, which is the the posterior probability characterizing one's best estimate of the true state of $m$ given noisy memory content $s$ and new evidence $a$, conditioned on whether the new evidence is plausible or not $z$. Figure 4 shows how $p(m|s,z,a)$ is dependent on anchor plausibility. Intuitively, when the anchor is considered to be plausible, $p(m|s,z=1,a)$, memory is biased toward the anchor, which in this case we fixed at $a=120$; when the anchor is considered to be implausible, $p(m|s,z=0,a)$ is centered around the noisy memory content $s=60$ and not biased by the anchor (we took $s=60$ for this example, assuming the memory trace is veridical). This effect is evident from Equations 7 and 8. Given $\sigma_0^2 >> \sigma_1^2$, this leads to the anchor $a$ being weighted more when the anchor is plausible ($z=1$), whereas $a$ being weighted less when the anchor is not plausible ($z=0$).

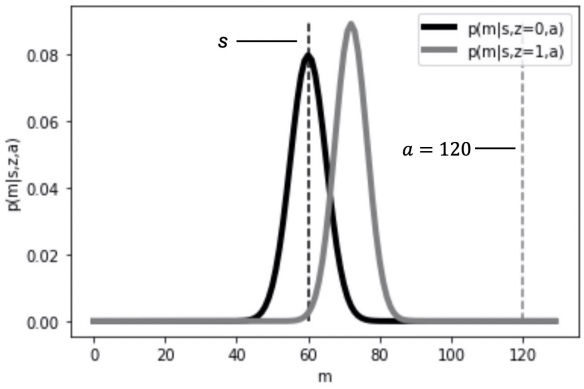


Figure 4: Simulation of $p(m|s,z,a)$ for a fixed anchor value, $a=120$. Dotted lines represent the mean of the memory, $s$, and the anchor value, $a$.

For each anchor value, we can obtain $p(z=0|a,s)$ and $p(z=1|a,s)$. Figure 5 shows the distribution of $p(z|a,s)$ for different values of the anchor, $a$. Consistent with Equation 10 and 11, when the anchor is farther from the noisy memory, $s$, it is more likely to be implausible, which explains the high values at the edges for the black curve, $p(z=0|a,s)$ and the low values near the center. Similarly, when the anchor is closer to the noisy memory, $s$, it is more likely to be inferred as plausible, $z=1$, explaining the high values in the center and low values near the edges of the gray curve.

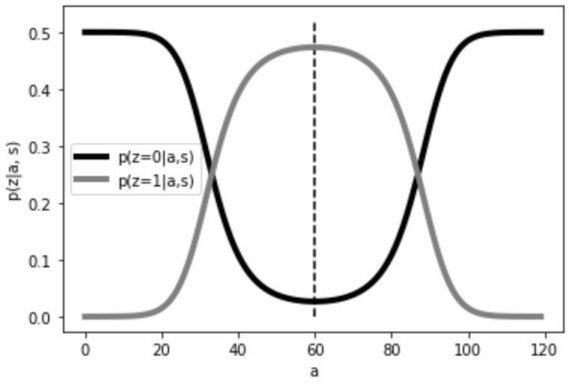


Figure 5: Simulation of $p(z|a,s)$ for values of a. While the anchor is implausible, it is more likely to have a value far from the mean of the expectation. When the anchor is plausible, it is more likely to be close to this mean.

Because our posterior distribution is a Gaussian mixture model with different means and variances, $E(m|a,s)$ is simply the mean of the distributions $p(m|s,z,a)$ weighted by the probabilities $p(z|a,s)$. We calculated the analytical mean by computing the mean of each category, $p(m|s,z=0,a)$ and

$p(m|s,z=1,a)$, and averaging by $p(z=0|a,s)$ and $p(z=1|a,s)$ respectively.

To create the final curves, the values of hindsight bias, $E(m|a,s)-m$, were graphed against the magnitude of the anchor distance, $|a-m|$ in Figure 6. A graph of hindsight bias versus anchor values is also included in Figure 7. Figure 6 provides a close match to the desired curve in Figure 1, capturing the inverse U-shape relationship between anchor distance and hindsight bias. The model does not show a difference between negatively directed anchors and positively directed anchors as seen from the experiment in Figure 1. We will discuss this discrepancy as part of the future work.

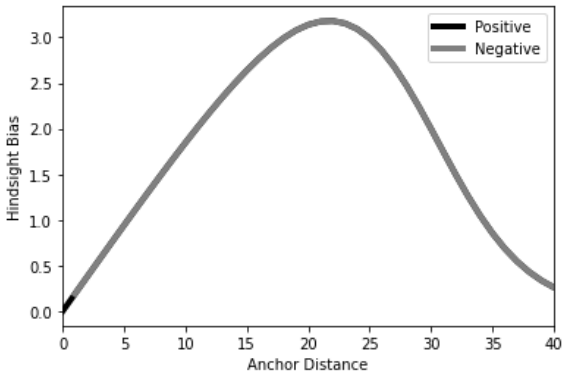


Figure 6: Hindsight bias ($\Delta h = E(m|a,s)-m$) and its dependence on anchor distance ($|a-m|$) as predicted by the Bayesian model for $m=s=60$. The model captures the inverse U-shape relationship between anchor distance and hindsight bias in Figure 1.

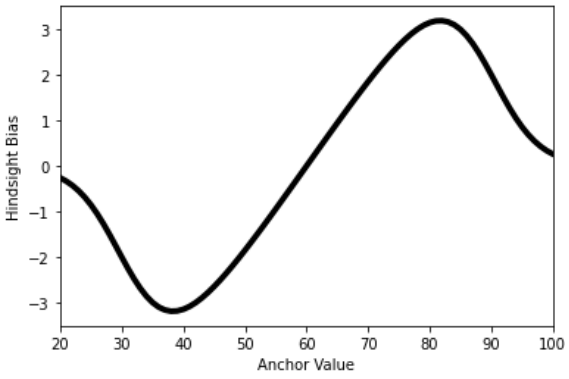


Figure 7: Hindsight bias and its dependence on specific anchor values as predicted by the Bayesian model. Note that the hindsight bias is negative when the anchor is less than $m$, positive when the anchor is greater than $m$, and zero when equal to $m$.

## Discussion

The results presented in this paper establish that the effect of anchor plausibility on hindsight bias can be explained as the consequence of optimally solving the statistical problem of memory reconstruction using additional knowledge from the observed anchor. Hardt and Pohl (2003) hypothesized that the U-shaped relationship between anchor distance and amount of hindsight bias is the result of two independent processes. More precisely, the impact of the anchor increases with anchor distance, but the probability of a biased reconstruction depends on anchor plausibility. However, we proposed that the impact from anchor distance and the impact from anchor plausibility are not two independent processes; our computational model captures how they jointly affect the amount of hindsight bias. The amount of hindsight bias exhibited depends on both anchor distance and the extent the anchor is perceived to be plausible.

The current work closely relates to the existing literature on modeling the effects of category structure on perception. In the model presented by Huttenlocher, Hedges and Vevea (2000), people used category structure to compensate for uncertainty in memory of sizes; in Feldman and Griffiths's (2007) model, category structure is used to correct for uncertainty in speech signals. The key difference between these models and our own is that in our model, the inference of category is from the additional evidence of the anchor rather than from the original stimuli.

The proposed model simulates the qualitative patterns of the inverse U-shaped relationship between anchor distance and amount of hindsight bias. Hindsight bias increases as the anchor distance increases, as long as the anchor is considered to be plausible; However, as anchor distance keeps increasing, the anchor is considered to be more and more implausible, hindsight bias starts decreasing. One direction for future work is to evaluate the model's fit to individual subject data. Another direction for future work entails explaining the additional finding that hindsight bias was larger for negatively directed anchors than for positively directed anchors in Figure 1. Hardt and Pohl suggested that the direction effect is potentially contributed to by the distribution of the estimates. When the distribution is skewed, anchors in one direction will become unacceptable sooner than anchors with the same amount of deviation in the opposite direction (Hardt & Pohl, 2003). Further experimental and modeling research is needed to verify this hypothesis.


## Acknowledgements

This work was supported by a C.V. Starr Fellowship awarded to Q.Z.


## References


Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making, 7(4)*, 223-242.

Christensen-Szalankski, J. J. J. & Beach, L. R. (1984). The citation bias: Fad and Fashion in the judgement and

decision literature. *American Psychologist, 39*, 75-78.

Christensen-Szalanski, J. J. J. & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48(1), 147-168*.

Erdfelder, E., Brandt, M., & Bröder, A. (2007). Recollection biases in hindsight judgments. *Social Cognition, 25(1)*, 114–131.

Feldman, N. H.& Griffiths, T. L. (2007). A Rational account of the perceptual magnet effect. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Hardt, O., & Pohl, R. (2003). Hindsight bias as a function of anchor distance and anchor plausibility. *Memory, 11(4-5)*, 379-394.

Hasher, L., Attig, M. S., & Alba, J. W. (1981). I knew it all along: Or, did I? *Journal of Verbal Learning and Verbal Behavior, 20*, 86-96.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107(3)*, 311-327.

Heckerman, D. (1995). *A Tutorial on learning With Bayesian networks.*Technical Report, Microsoft Research Advanced Technology Division.

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory and Cognition, 16(6)*, 533-538.

Hemmer, P., & Steyvers, M. (2009), A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science, 1*, 189-202.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why Do Categories Affect Stimulus Judgement? Journal of *Experimental Psychology: General, 129(2)*, 220-241.

Lieder, F., Griffiths, T., Huys, Q. J., & Goodman, N. D. (2017). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin Review, 25*, 322-349.

Parpart, P., Jones, M., & Love, B.C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology, 102*, 127-144.

Pohl, R. F. (1998). The effects of feedback source and plausibility on hindsight bias. *European Journal of Cognitive Psychology, 10(2)*, 191-212.

Pohl, R., Eisenhauer, M., & Hardt, O. (2003). SARA: A cognitive process model to simulate the anchoring effect and hindsight bias. *Memory, 11(4-5)*, 337-356.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation, 14*, 207-262. San Diego, CA: Academic Press.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on psychological science, 7(5)*, 411-426.

Roese, N. J. (2012) "I Knew It All Along…Didn't I?" – Understanding Hindsight Bias. *APS Research News*.

Shiffrin, R. M., & Raaijmakers, J. G. W. (1992). The SAM retrieval model: A retrospective and a prospective. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in Honor of William K. Estes, 2*, 69-86. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition, 101*, 467-494.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology, 73(3)*, 437-446.

Synodinos, N. E. (1986). Hindsight distortion: "I-knew-it-all-along and I was sure about it". *Journal of Applied Social Psychology, 16*, 107-117.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, D.T. & Jarvis, W. B. G. (2001). Implications of Attitude Change Theories for Numerical Anchoring: Anchor Plausibility and the Limits of Anchor Effectiveness. *Journal of Experimental Social Psychology, 37(1)*, 62-69.