Psychon Bull Rev
DOI 10.3758/s13423-017-1286-8

THEORETICAL REVIEW

The anchoring bias reflects rational use of cognitive resources

Falk Lieder[1,2] · Thomas L. Griffiths[1,5] · Quentin J. M. Huys[2,4] · Noah D. Goodman[3]

Abstract Cognitive biases, such as the anchoring bias, pose a serious challenge to rational accounts of human cognition. We investigate whether rational theories can meet this challenge by taking into account the mind's bounded cognitive resources. We asked what reasoning under uncertainty would look like if people made rational use of their finite time and limited cognitive resources. To answer this question, we applied a mathematical theory of bounded rationality to the problem of numerical estimation. Our analysis led to a rational process model that can be interpreted in terms of anchoring-and-adjustment. This model provided a unifying explanation for ten anchoring phenomena including the differential effect of accuracy motivation on the bias towards provided versus self-generated anchors. Our results illustrate the potential of resource-rational analysis to provide formal theories that can unify a wide range of empirical results and reconcile the impressive capacities of the human mind with its apparently irrational cognitive biases.

Keywords Bounded rationality · Heuristics · Cognitive biases · Probabilistic reasoning · Anchoring-and-adjustment · Rational process models

Many classic theories in economics, philosophy, linguistics, social science, and psychology are built on the assumption that humans are rational (Friedman & Savage, 1948; Lohmann, 2008; Hedström & Stern, 2008; Harman, 2013; Frank & Goodman, 2012) and therefore act according to the maxims of expected utility theory (Von Neumann & Morgenstern, 1944) and reason according to the laws of logic (Mill, 1882; Newell et al., 1958; Braine, 1978; Fodor, 1975) or probability theory (Oaksford & Chater, 2007). The assumption that people are rational was challenged when a series of experiments suggested that people's judgments systematically violate the laws of logic (Wason, 1968) and probability theory (Tversky & Kahneman, 1974). For instance, Tversky and Kahneman (1974) showed that people's probability judgments appear to be insensitive to prior probability and sample size but are influenced by irrelevant factors such as the ease of imagining an event or the provision of an unrelated random number. These systematic deviations from the tenets of logic and probability are known as *cognitive biases*. According to Tversky and Kahneman (1974), cognitive biases result from people's use of fast but fallible cognitive strategies known as *heuristics*.

✉ Falk Lieder
falk.lieder@berkeley.edu

Thomas L. Griffiths
tom_griffiths@berkeley.edu

Quentin J. M. Huys
qhuys@biomed.ee.ethz.ch

Noah D. Goodman
ngoodman@stanford.edu

[1] Helen Wills Neuroscience Institute, University of California, Berkeley, USA

[2] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zürich and Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

[3] Department of Psychology, Stanford University, Stanford, USA

[4] Department of Psychiatry, Psychotherapy and Psychosomatics, Hospital of Psychiatry, University of Zürich, Zürich, Switzerland

[5] Department of Psychology, University of California, Berkeley, USA

Published online: 08 May 2017

The discovery of cognitive biases was influential because following the rules of logic and probability was assumed to be the essence of rational thinking. Evidence that people deviate from these rules brings human rationality into question. This doubt is shaking the foundations of economics, the social sciences, and rational models of cognition. If the human mind does not follow rational principles, then there is little hope that we will be able to able derive unifying laws of cognition from a basic set of axioms. Without the principles of rationality, there is little guidance for how to translate assumptions about cognitive processes into predictions about behavior and how to generalize from our data. But if people were systematically rational in some sense then all of this would be possible, and creating artificial intelligence could go hand in hand with understanding how the mind works. Therefore, the question whether people are rational is fundamental to how we study the mind, to how we model it, and the implications of our theories for science and society.

Despite their cognitive biases, humans still outperform intelligent systems built on the laws of logic and probability on many real-world problems. This poses a paradox: how can we be so smart, if we appear so irrational? The argument that people are irrational rests on two premises: First, to be rational is to follow the rules of logic and probability theory. Second, human thought violates the rules of logic and probability. Previous work supports the second premise (Shafir & LeBoeuf, 2002), but in this article we question the first by suggesting that notions of human rationality should take into account that reasoning costs time. The number of computations required for exact logical or probabilistic reasoning grows exponentially with the number of facts and variables to be considered. As a result, to exactly and completely reason through just a single complex everyday situation involving hundreds of variables could require more computations than can be performed in a human lifetime (Van Rooij, 2008). Thus, if a person were to reason out everything strictly according to the laws of logic and probability theory she might die before she reached her first conclusion.

The laws of logic and probability theory are thus insufficient to give a definition of rationality relevant to any real intelligent agent, because the cost of computation has to be taken into account. To be successful in the world we live in, we have to solve complex problems in finite time despite bounded cognitive resources. In this paper, we explore the implications of a different framework for characterizing rationality that captures this idea: *resource-rationality* (Lieder et al., 2012; Griffiths et al., 2015), which builds on the notion of *bounded optimality* proposed in the artificial intelligence literature by Russell and colleagues (Russell & Wefald, 1991; Russell & Subramanian, 1995; Russell, 1997). We use this alternative characterization of rationality to re-evaluate human performance in tasks used to demonstrate that people's judgments are biased because they are cognitive misers. Achieving demanding goals in limited time requires balancing being quick and being accurate. We regret the opportunities we miss when we fail to make up our mind on time, but we also regret the errors we commit by jumping to conclusions. When we think too little our judgments can be skewed by irrelevant information that we happened to see, hear, or think about a moment ago. This phenomenon is known as *anchoring*. Anchoring is one of the cognitive biases discovered by Tversky and Kahneman (1974). It impacts many important aspects of our lives including the outcome of salary negotiations (Galinsky & Mussweiler, 2001), economic decisions (e.g., Simonson & Drolet, 2004), criminal sentences (Englich et al., 2006), and even our ability to understand other people (Epley et al., 2004).

In their classic paper, Tversky and Kahneman (1974) showed that people's judgments could be systematically skewed by providing them with an arbitrary number before their judgment: The experimenter generated a random number by spinning a wheel of fortune, and then asked participants to judge whether the percentage of African countries in the United Nations was smaller or larger than that number. Participants were then asked to estimate this unknown quantity. Strikingly, the participants' estimates were biased towards the random number: their median estimate was larger when the random number was high than when it was low. This appears to be a clear violation of rationality. According to Tversky and Kahneman (1974) this violation occurs because people use a two-stage process called *anchoring-and-adjustment* (see also Nisbett & Ross, 1980). In the first stage, people generate a preliminary judgment called their *anchor*. In the second stage, they adjust that judgment to incorporate additional information, but the adjustment is usually insufficient. In Tversky and Kahneman's experiment people appear to have anchored on the random number provided by the experimenter and adjusted it insufficiently. Consequently, when the anchor was low people's judgments were too low, and when the anchor was high their judgments were too high.

At first sight, anchoring appears to be irrational, because it deviates from the standards of logic and probability which are typically used to assess rationality. But it could also be a reasonable compromise between error in judgment and the cost of computation, and hence be resource-rational. Anchoring-and-adjustment has two components that could be irrational: the generation of the anchor and the process by which it is adjusted. Previous research found that when no anchor is provided, the anchors that people generate for themselves are relevant quantities that are reasonably close to the correct value and can be generated quickly (Epley & Gilovich, 2006). Furthermore, research on human communication suggests that in everyday life it is reasonable to assume that other people are cooperative and provide relevant information (Schwarz, 2014). Applied to anchoring, this means that if somebody asks you in real life whether a quantity you know very little about is larger or smaller than a certain value, it would be rational to treat that question as a clue to its value (Zhang & Schwarz, 2013). Thus, having the queried value in mind might make it rational to reuse it as your anchor for estimating the unknown quantity. This suggests that the mechanism by which people generate their anchors could be rational in the real world.[1]

Assuming that people generate or select anchors in a reasonable way, the rationality of anchoring-and-adjustment hinges on the question whether adjustment is a rational process. To answer this question, we investigate whether insufficient adjustment can be understood as a rational tradeoff between time and accuracy. If so, then how much people adjust their initial estimate should adapt rationally to the relative utility of being fast versus being accurate. To formalize this hypothesis, we present a resource-rational analysis of numerical estimation. Our analysis suggests that the rational use of finite resources correctly predicts the anchoring bias and how it changes with various experimental manipulations (see Table 1). These results support the conclusion that adjustment is resource-rational.

The remainder of this article begins with a brief survey of empirical findings on anchoring and discusses the challenges that they pose to existing accounts of anchoring-and-adjustment. We then present our resource-rational analysis of numerical estimation, derive a rational process model that can be interpreted in terms of anchoring-and-adjustment, and show it is sufficient to explain the reviewed phenomena. We close by discussing our findings and their implications for the debate about human rationality.

Empirical findings on the anchoring bias

Anchoring is typically studied in numerical estimation tasks. Numerical estimation involves making an informed guess of the value of an unknown numerical quantity. Since the first anchoring experiment by Tversky and Kahneman (1974) a substantial number of studies have investigated when anchoring occurs and what determines the magnitude of the anchoring bias (see Table 1).

The anchors that people use when forming estimates can be relevant to the quantity they are estimating. For instance, Tversky and Kahneman (1974) found that people sometimes anchor on the result of calculating $1 \times 2 \times 3 \times 4$ when the task is estimating $1 \times 2 \times 3 \times 4 \times \cdots \times 8$. However, people can also be misled, anchoring on numbers that are irrelevant to the subsequent judgment. For instance, many anchoring experiments first ask their participants whether an unknown quantity is larger or smaller than a given value and then proceed to have them estimate that quantity. Having compared the unknown quantity to the value provided by the experimenter makes people re-use that value as their anchor in the subsequent estimation task. Those numbers are therefore known as *provided anchors*. Importantly this procedure works with irrelevant numbers such as the random number that Tversky and Kahneman (1974) generated for their participants or one's own social security number (Ariely et al., 2003).

Although asking people to compare the quantity to a given number is particularly effective, the anchoring bias also occurs when anchors are presented incidentally (Wilson et al., 1996), although this effect is smaller and depends on particulars of the anchor and its presentation (Brewer & Chapman, 2002). Furthermore, anchoring-and-adjustment can also occur without an externally provided anchor: At least in some cases people appear to generate their own anchor and adjust from it Epley and Gilovich (2004). For instance, when Americans are asked to estimate the boiling point of water on Mount Everest they often recall 212 °F (100 °C) and adjust downwards to accommodate the lower air pressure in higher altitudes.

Although people's adjustments are usually insufficient, various factors influence their size and consequently the magnitude of the anchoring bias. For instance, the anchoring bias is larger the more uncertain people are about the quantity to be estimated (Jacowitz & Kahneman, 1995). Indeed, Wilson et al. (1996) found that people knowledgeable about the quantity to be estimated were immune to the anchoring bias whereas less knowledgeable people were susceptible to it. While familiarity (Wright and Anderson, 1989) and expertise (Northcraft & Neale, 1987) do not abolish anchoring, expertise appears to at least reduce it Northcraft and Neale (1987). Other experiments have systematically varied the distance from the anchor to the correct value. Their results suggested that the magnitude of the anchoring bias initially increases with the distance from the anchor to the correct value (Russo & Schoemaker, 1989). Yet this linear in crease of the anchoring bias does not continue indefinitely. Chapman and Johnson (1994) found that increasing an already unrealistically large anchor increases the anchoring bias less than increasing a realistic anchor by the same amount.

Critically for the resource-rational account proposed here, the computational resources available to people also seem to influence their answers. Time pressure, cognitive load, and alcohol decrease the size of people's adjustments and inter-individual differences in how much people adjust their initial estimate correlate with relevant personality traits such as the need for cognition (Epley and Gilovich, 2006). In addition to effects related to cognitive resources,

[1] We will revisit this issue in more depth in the general discussion.

Table 1 Anchoring phenomena and resource-rational explanations

Anchoring effect	Simulated results	Resource-rational explanation
Insufficient adjustment from provided anchors	Jacowitz and Kahneman (1974), Tversky and Kahneman (1995)	Rational speed-accuracy tradeoff.
Insufficient adjustment from self-generated anchors	Epley and Gilovich (2006), Study 1	Rational speed-accuracy tradeoff.
Cognitive load, time pressure, and alcohol reduce adjustment.	Epley and Gilovich (2006), Study 2	Increased cost of adjustment reduces the resource-rational number of adjustments.
Anchoring bias increases with anchor extremity.	Russo and Schoemaker (1989)	Each adjustment reduces the bias by a constant factor (3). Since the resource-rational number of adjustments is insufficient, the bias is proportional to the distance from anchor to correct value.
Uncertainty increases anchoring.	Jacowitz and Kahneman (1995)	The expected change per adjustment is small when nearby values have similar plausibility.
Knowledge can reduce the anchoring bias.	Wilson et al. (1996), Study 1	High knowledge means low uncertainty. Low uncertainty leads to high adjustment (see above).
Accuracy motivation reduces anchoring bias when the anchor is self-generated but not when it is provided.	Tversky and Kahneman (1974), Epley and Gilovich (2005)	1. People are less uncertain about the quantities for which they generate their own anchors. 2. Accuracy motivation increases the number of adjustments but change per adjustment is lower when people are uncertain.
Telling people whether the correct value is larger or smaller than the anchor makes financial incentives more effective.	Simmons et al. (2010), Study 2	Being told the direction of adjustments makes adjustments more effective, because adjustments in the wrong direction will almost always be rejected.
Financial incentives are more effective when the anchor is extreme.	Simmons et al. (2010), Study 3	Values on the wrong side of an extreme anchor are much less plausible than values on the correct side. Therefore proposed adjustments in the wrong direction will almost always be rejected.

adjustment also depends on incentives. Intuitively, accuracy motivation should increase the size of people's adjustments and therefore decrease the anchoring bias. Interestingly, experiments have found that accuracy motivation decreases the anchoring bias only in some cases, but not in others (Epley & Gilovich, 2006; Simmons et al., 2010). On questions where people generated their own anchors, financial incentives increased adjustment and reduced the anchoring bias (Epley & Gilovich, 2006; Simmons et al., 2010). But on questions with provided anchors, financial incentives have typically failed to eliminate or reduce the anchoring bias (Tversky & Kahneman, 1974; Ariely et al., 2003) with some exceptions (Wright & Anderson, 1989). A recent set of experiments by Simmons et al. (2010) suggested that accuracy motivation increases adjustment from provided and self-generated anchors if and only if people know in which direction to adjust. Taken together, these findings suggests that the anchoring bias depends on how much cognitive resources people are able to and willing to invest.

Before the experiments by Simmons et al. (2010) demonstrated that accuracy motivation can increase adjustment from provided anchors, the bias towards provided anchors appeared immutable by financial incentives (Tversky and Kahneman, 1974; Chapman & Johnson, 2002; Wilson et al., 1996), forewarnings and time pressure (Mussweiler & Strack, 1999; but see Wright & Anderson, 1989). Since incentives were assumed to increase adjustment and increased adjustment should reduce the anchoring bias, the ineffectiveness of incentives led to the conclusion that the anchoring bias results from a mechanism other than anchoring-and-adjustment, such as selective accessibility (Mussweiler & Strack, 1999; Chapman & Johnson, 2002; Epley, 2004). Later experiments found that when people generate the anchor themselves accuracy motivation and time pressure are effective (Epley & Gilovich, 2005; Epley et al., 2004; Epley & Gilovich, 2006). This led (Epley & Gilovich, 2006) to conclude that people use the anchoring-and-adjustment strategy only when they generated the anchor themselves whereas provided anchors bias judgments through a different mechanism.

The wide range of empirical phenomena summarized in Table 1 have suggested a correspondingly wide range of explanations, including the idea that anchoring and adjustment is not a simple, unitary process. In the remainder of the paper we explore an alternative account, showing that these disparate and seemingly inconsistent phenomena can all be explained by a unifying principle: the rational use of finite time and cognitive resources. From this principle we derive a resource-rational anchoring-and-adjustment model and show that it is sufficient to explain the anchoring bias regardless of whether the anchor was provided or self-generated.

Anchoring and adjustment as resource-rational inference

In this section we formalize the problem people solve in anchoring experiments – numerical estimation – and analyze how it can be efficiently solved in finite time with bounded cognitive resources. We thereby derive a resource-rational model of anchoring-and-adjustment. We then use this model to explain a wide range of anchoring phenomena.

Conceptually, our model assumes that adjustment proceeds by repeatedly considering small changes to the current estimate. The proposed change is accepted or rejected probabilistically such that the change is more likely to be made the more probable the new value is and the less probable the current one is (see Fig. 1). After sufficiently many adjustments the estimate becomes correct on average and independent of the initial guess. However, each small adjustment costs a certain amount of time. According to our model, the number of steps is chosen to minimize the expected value of the time cost of adjustment plus the error cost of the resulting estimate. In the remainder of this section, we derive our model from first principles, specify it in detail, and show that the optimal number of adjustments is very small. As Fig. 1 illustrates, this causes the final estimates to be biased towards their respective anchors.

In contrast to previous theories of anchoring (Epley & Gilovich, 2006; Simmons et al., 2010), our model precisely specifies the number, size, and direction of adjustments as a function of the task's incentives and the participant's knowledge. In contrast, to the proposal by Epley and Gilovich (2006) our model covers adjustments from provided anchors *and* self-generated anchors. Furthermore, while Epley and Gilovich (2006) assumed that the correct direction of adjustment is known, our model does not make this assumption and allows the direction of adjustment to change from one step to the next. The model by Simmons et al. (2010) also makes these conceptual assumptions. However, it does not specify precisely how the direction and size of each adjustment are determined. While their model predicts a deterministic back-and-forth in the face of uncertainty, our model assumes that adjustments that improve the estimate are probabilistically preferred to adjustments that do not. This enables our model to capture streaks of adjustments in the correct direction interrupted by small steps in the wrong direction, whereas the model by Simmons et al. (2010) appears to predict that the direction of adjustment should constantly alternate. Finally, while both previous models assumed that adjustment stops as soon as the current estimate is sufficiently plausible (Epley and Gilovich, 2006; Simmons et al., 2010), we propose that the number of adjustments is pre-determined adaptively to achieve an optimal speed-accuracy tradeoff on average. In the subsequent section we apply the resulting model to explain the various anchoring phenomena summarized in Table 1.

Resource-rational analysis

Resource-rational analysis is a new approach to answering a classic question: how should we think and decide given that our time and our minds are finite? In economics this problem was first identified by Simon (1955, 1956, 1972). Simon pointed out that our finite computational capacities make it impossible for us to always find the best course of action, because we cannot consider all possible consequences. He illustrated this using the game of chess, where choosing the optimal move would require considering about 10^{120} possible continuations. Thus, Simon concluded, to adequately model human behavior we need a theory of rationality that takes our minds' limits into account. Simon called such an approach *bounded rationality*, emphasizing that it depends on the structure of the environment (Simon, 1956) and entails satisficing, that is accepting suboptimal solutions that are good enough, rather than optimizing. Subsequent research has identified simple heuristics that make good, but not necessarily optimal, decisions and judgments (Tversky 1972; Thorngate 1980; McKenzie 1994; Gigerenzer & Goldstein 1996) very efficiently. Thinking is assumed to be costly and alternative strategies differ in the amount of thinking they entail (e.g., Shugan, 1980). Based on this line of reasoning, it has been proposed that people adaptively select their cognitive strategies from a toolbox of simple heuristics (Gigerenzer & Selten, 2002) according to a cost-benefit analysis (Beach & Mitchell, 1978; Payne et al., 1993).

While Simon (1955) provided some formal examples of satisficing strategies, he viewed bounded optimality as a principle rather than a formal framework. Subsequent researchers have tried to formally capture the tradeoff between time and errors. Good (1983) formulated this idea in terms of the maximization of expected utility taking into account deliberation cost. Intuitively, this means that rational bounded agents optimally trade off the expected utility of the action that will be chosen with the corresponding deliberation cost. Yet, Good (1983) did not make this notion mathematically precise. Furthermore, his formulation does

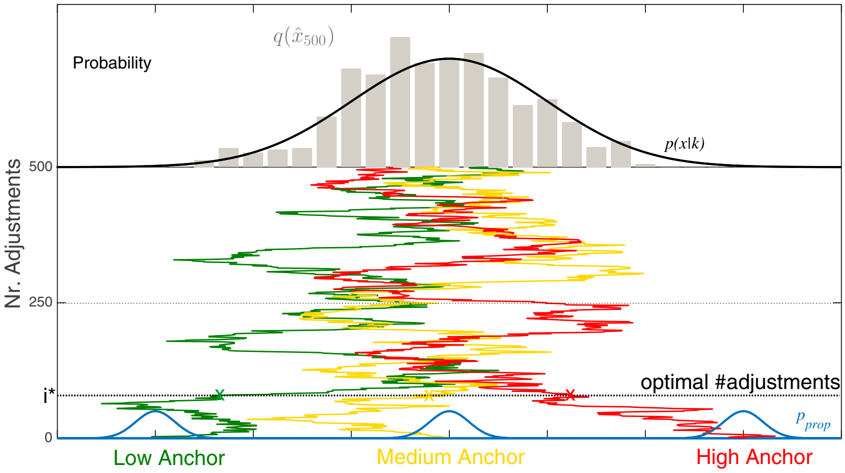

Fig. 1 The figure illustrates the resource-rational anchoring-and-adjustment. The three jagged lines are examples of the stochastic sequences of estimates the adjustment process might generate starting from a low, medium, and high anchor respectively. In each iteration a potential adjustment is sampled from a proposal distribution p_{prop} illustrated by the bell curves. Each proposed adjustment is stochastically accepted or rejected such that over time the relative frequency with which different estimates are considered $q(\hat{x}_t)$ becomes the target distribution $p(x|k)$. The top of the figure compares the empirical distribution of the samples collected over the second half of the adjustments with the target distribution $p(x|k)$. Importantly, this distribution is the same for each of the three sequences. In fact, it is independent of the anchor, because the influence of the anchor vanishes as the number of adjustments increases. Yet, when the number of adjustments (iterations) is low (e.g., 25), the estimates are still biased towards their initial values. The optimal number of iterations $i^{\star}$ is very low as illustrated by the dotted line. Consequently, the resulting estimates indicated by the red, yellow, and red cross are still biased towards their respective anchors

not take into account the deliberation cost of determining the optimal tradeoff between expected utility and deliberation cost. These problems were solved by Russell and colleagues (Russell and Wefald, 1991; Russell & Subramanian, 1995; Russell, 1997) who provided a complete, formal, mathematical theory of the rationality of bounded agents. In this framework, agents are considered to be rational if they follow the *algorithm* that makes the best possible use of their computational architecture (e.g., hardware) and time.

Resource-rational analysis leverages this abstract theory for understanding the human mind. To be resource-rational is to make optimal use of one's finite time and limited cognitive resources. Resource-rational analysis (Griffiths et al., 2015) derives rational process models of cognitive abilities from formal definitions of their function and abstract assumptions about the mind's computational architecture. This function-first approach starts at the computational level of analysis (Marr, 1982). When the problem solved by the cognitive capacity under study has been formalized, resource-rational analysis postulates an abstract computational architecture, that is a set of elementary operations and their costs, with which the mind might solve this problem. Next, resource-rational analysis derives the algorithm that is optimal for solving the problem identified at the computational level with the abstract computational architecture. The resulting process model can be used to simulate people's responses and reaction times in a given experiment. The model's predictions are tested against empirical data. Based on this evaluation, the assumptions about the computational architecture and the problem to be solved are revised.

Resource-rational analysis of numerical estimation

Having introduced the basic concepts of resource rationality, we now apply resource-rational analysis to numerical estimation: We start by formalizing the problem solved by numerical estimation. Next, we specify an abstract computational architecture. We then derive the optimal solution to the numerical estimation problem afforded by the computational architecture. This resource-rational strategy will then be evaluated against empirical data in the remainder of this article.

Function In numerical estimation people have to make an informed guess about an unknown quantity X based on their knowledge K. In general, people's relevant knowledge K is incomplete and insufficient to determine the quantity X with certainty. For instance, people asked to estimate the boiling point of water on Mount Everest typically do not know its exact value, but they do know related information, such as the boiling point of water at normal altitude, the freezing point of water, the qualitative relationship between altitude, air pressure, and boiling point, and so on. We formalize people's uncertain belief about X by the probability distribution $P(X|K)$ which assigns a plausibility $p(X = x|K)$ to each potential value x. According to Bayesian decision theory, the goal is to report the estimate $\hat{x}$ with the highest expected utility $\mathbb{E}_{P(X|K)}[u(\hat{x}, x)]$. This is equivalent to finding the estimate with the lowest expected error cost

$$x^\star = \arg\min_{\hat{x}} \mathbb{E}_{P(X|K)}[\text{cost}(\hat{x}, x)], \tag{1}$$

where $x^\star$ is the optimal estimate, and $\text{cost}(\hat{x}, x)$ is the error cost of the estimate $\hat{x}$ when the true value is x. Here, we assume that the error cost is the absolute deviation of the estimate from the true value, that is $\text{cost}(\hat{x}, x) = |\hat{x} - x|$.

Model of mental computation How the mind should solve the problem of numerical estimation (see Eq. 1) depends on its computational architecture. Thus, to derive predictions from the assumption of resource-rationality we have to specify the mind's elementary operations and their cost. To do so, we build on the resource-rational analysis by Vul et al. (2014) which assumed that the mind's elementary computation is *sampling*. Sampling is widely used to solve inference problems in statistics, machine learning, and artificial intelligence (Gilks et al., 1996). Several behavioral and neuroscientific experiments suggest that the brain uses computational mechanisms similar to sampling for a wide range of inference problems ranging from vision to causal learning (Vul et al., 2014; Denison et al., 2013; Bonawitz et al., 2014; Bonawitz et al., 2014; Griffiths and Tenenbaum, 2006; Stewart et al., 2006; Fiser et al., 2010). One piece of evidence is that people's estimates of everyday events are highly variable even though the average of their predictions tends to be very close to the optimal estimate prescribed by Bayesian decision theory (see Eq. 1, Griffiths & Tenenbaum, 2006, 2011). Furthermore, Vul et al. (2014) found that the relative frequency with which people report a certain value as their estimate is roughly equal to its posterior probability, as if the mind was drawing one sample from the posterior distribution.

Sampling stochastically simulates the outcome of an event or the value of a quantity such that, on average, the relative frequency with which each value occurs is equal to its probability. According to Vul et al. (2014), people may estimate the value of an unknown quantity X using only a single sample from the subjective probability distribution $P(X|K)$ that expresses their beliefs. If the expected error cost (1) is approximated using a single sample $\tilde{x}$, then that sample becomes the optimal estimate. Thus, the observation that people report estimates with frequency proportional to their probability is consistent with them approximating the optimal estimate using only a single sample.

However, for the complex inference problems that people face in everyday life generating even a single perfect sample can be computationally intractable. Thus, while sampling is a first step from computational level theories based on probabilistic inference towards cognitive mechanisms, a more detailed process model is needed to explain how simple cognitive mechanisms can solve the complex inference problems of everyday cognition. Here, we therefore explore a more fine-grained model of mental computation whose elementary operations serve to approximate sampling. In statistics, machine learning, and artificial intelligence sampling is often approximated by Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1996). MCMC algorithms allow the drawing of samples from arbitrarily complex distributions using a stochastic sequence of approximate samples, each of which depends only on the previous one. Such stochastic sequences are called Markov chains; hence the name Markov chain Monte Carlo.

In the remainder of the paper, we explore the consequences of assuming that people answer numerical estimation questions by engaging in a thought process similar to MCMC. We assume that the mind's computational architecture supports MCMC by two basic operations: The first operation takes in the current estimate and stochastically modifies it to generate a new one. The second operation compares the posterior probability of the new estimate to that of the old one and accepts or rejects the modification stochastically. Furthermore, we assume that the cost of computation is proportional to how many such operations have been performed. These two basic operations are sufficient to execute an effective MCMC strategy for probabilistic inference known as the Metropolis-Hastings algorithm (Hastings, 1970). This algorithm is the basis for our anchoring-and-adjustment models as illustrated in Fig. 1.

To be concrete, given an initial guess $\hat{x}_0$, which we will assume to be the anchor a ($\hat{x}_0 = a$), this algorithm performs a series of adjustments. In each step a potential adjustment δ is proposed by sampling from a symmetric probability distribution P_{prop} ($\delta \sim P_{\text{prop}}, P_{\text{prop}}(-\delta) = P_{\text{prop}}(\delta)$). The adjustment will either be accepted, that is $\hat{x}_{t+1} = \hat{x}_t + \delta$, or rejected, that is $x_{t+1} = \hat{x}_t$. If a proposed adjustment makes the estimate more probable ($P(X = \hat{x}_t + \delta|K) > P(X = \hat{x}_t|K)$), then it will always be accepted. Otherwise the adjustment will be made with probability $\alpha = \frac{P(X=\hat{x}_t+\delta|K)}{P(X=\hat{x}_t|K)}$, that is according to the posterior probability of the adjusted relative to the unadjusted estimate. This strategy ensures that regardless of which initial value you start from, the frequency with which each value x has been considered will eventually equal to its subjective probability of being correct, that is $P(X = x|K)$. This is necessary to capture the finding that the distribution of people's estimates is very similar to the posterior distribution $P(X = x|K)$ (Vul et al., 2014; Griffiths and Tenenbaum, 2006). More formally, we can say that as the number of adjustments t increases, the distribution of estimates $Q(\hat{x}_t)$ converges to the posterior distribution $P(X|K)$. This model of computation has the property that each adjustment decreases an upper bound on the expected error by a constant multiple (Mengersen & Tweedie, 1996). This property is known as geometric convergence and illustrated in Fig. 2.

There are several good reasons to consider this computational architecture as a model of mental computation in the domain of numerical estimation: First, the success of MCMC methods in statistics, machine learning, and artificial intelligence suggests they are well suited for the complex inference problems people face in everyday life. Second, MCMC can explain important aspects of cognitive phenomena ranging from category learning (Sanborn et al., 2010) to the temporal dynamics of multistable perception (Moreno-Bote et al., 2011; Gershman et al., 2012), causal reasoning in children (Bonawitz et al., 2014), and developmental changes in cognition (Bonawitz et al., 2014). Third, MCMC is biologically plausible in that it can be efficiently implemented in recurrent networks of biologically plausible spiking neurons (Buesing et al., 2011). Last but not least, process models based on MCMC might be able to explain why people's estimates are both highly variable (Vul et al., 2014) and systematically biased (Tversky & Kahneman, 1974).

Optimal resource-allocation Resource-rational anchoring-and-adjustment makes three critical assumptions: First, the estimation process is a sequence of adjustments such that after sufficiently many steps the estimate will be a representative sample from the belief $P(X|K)$ about the unknown quantity X given the knowledge K. Second, each adjustment costs a fixed amount of time. Third, the number of adjustments is chosen to achieve an optimal speed-accuracy tradeoff. It follows, that people should perform the optimal number of adjustments, that is

$$t^\star = \arg\min_t \left[\mathbb{E}_{Q(\hat{x}_t)}\left[\text{cost}(x, \hat{x}_t) + \gamma \cdot t\right]\right], \quad (2)$$

where $Q(\hat{x}_t)$ is the distribution of the estimate after t adjustments, x is its unknown true value, $\hat{x}_t$ is the estimate after performing t adjustments, $\text{cost}(x, \hat{x}_t)$ is its error cost, and γ is the time cost per adjustment.

Figure 3 illustrates this equation showing how the expected error cost – which decays geometrically with the number of adjustments–and the time cost – which increases linearly – determine the optimal speed-accuracy tradeoff. We inspected the solution to Eq. 2 when the belief and the proposal distribution are standard normal distributions (i.e. $P(X|K) = P(X^{\text{prop}}) = \mathcal{N}(0, 1)$) for different anchors. We found that for a wide range of realistic time costs the optimal number of adjustments (see Fig. 4a) is much smaller than the number of adjustments that would be required to eliminate the bias towards the anchor. Consequently, the estimate obtained after the optimal number of adjustments is still biased towards the anchor as shown in the Fig. 4b. This is a consequence of the geometric convergence of the error (see Fig. 2) which leads to quickly diminishing returns for additional adjustments. This is a general property of this rational model of adjustment that can be derived mathematically (Lieder et al., 2012).

The optimal speed-accuracy tradeoff weights the costs in different estimation problems according to their prevalence in the agent's environment; for more information please see Appendix B.

Resource-rational explanations of anchoring phenomena

Following the definition of the bias of an estimator in mathematical statistics, we quantify the anchoring bias by $B_t(x, a) = \mathbb{E}[\hat{x}_t|x, a] - x$, where $\hat{x}_t$ is a participant's estimate of a quantity x after i adjustments, and a denotes the anchor. Figure 5 illustrates this definition and four basic ideas: First, the average estimate generated by anchoring-and-adjustment equals the anchor plus the adjustment. Second, the adjustment equals the relative adjustment times the total distance from the anchor to the posterior expectation. Third, adjustments tend to be insufficient, because the relative adjustment size is less than one. Therefore, the average

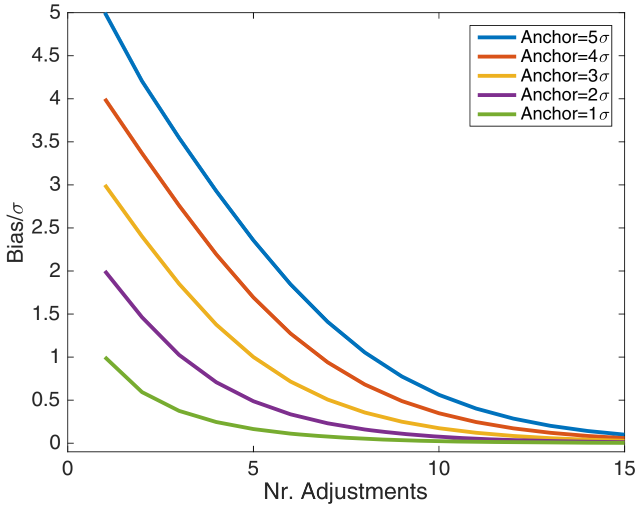

Fig. 2 In resource-rational anchoring-and-adjustment the bias of the estimate is bounded by a geometrically decaying function of the number of adjustments. The plots shows the bias of resource-rational anchoring-and-adjustment as a function of the number of adjustments for five different initial values located $1, \cdots, 5$ posterior standard deviations (i.e., σ) away from the posterior mean. The standard normal distribution was used as both the posterior $P(X|K)$ and the proposal distribution $P_{\text{prop}}(\delta)$

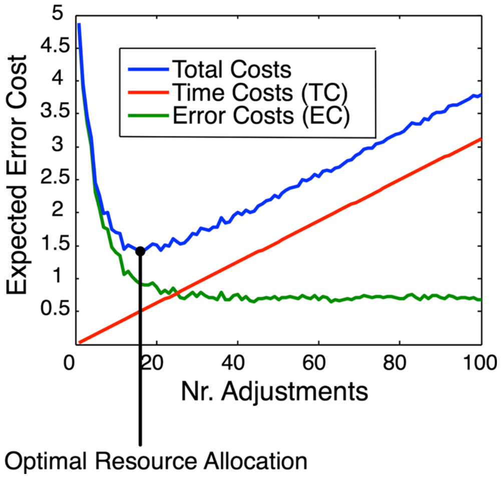

Fig. 3 The expected value of the error cost cost$(x, \hat{x}_n)$ shown in green decays nearly geometrically with the number of adjustments n. While the decrease of the error cost diminishes with the number of adjustments, the time cost $\gamma \cdot t$ shown in red continues to increase at the same rate. Consequently, there is a point when further decreasing the expected error cost by additional adjustments no longer offsets their time cost so that the total cost shown in blue starts to increase. That point is the optimal number of adjustments $t^\star$

estimate usually lies between the anchor and the correct value. Fourth, because the relative adjustment is less than one, the anchoring bias increases linearly with the distance from the anchor to the correct value.

More formally, the bias of resource-rational anchoring-and-adjustment cannot exceed a geometrically decaying function of the number of adjustments as illustrated in Fig. 2:

$$B_t(x, a) = \mathbb{E}[\hat{x}_t|x, a] - x \leq B_0(x, a) \cdot r^t = (a - x) \cdot r^t, \quad (3)$$

where r is the rate of convergence to the distribution $P(X|K)$ that formalizes people's beliefs. Consequently, assuming that the bound is tight, resource-rational anchoring-and-adjustment predicts that, on average, people's predictions $\hat{x}$ are a linear function of the correct value x and the anchor a:

$$\mathbb{E}[\hat{x}_t|x, a] \approx a \cdot r^t + (1 - r^t) \cdot x. \quad (4)$$

Therefore the anchoring bias remaining after a fixed number of adjustments increases linearly with the distance from the anchor to the correct value as illustrated in Fig. 5.

The hypothesis that the mind performs probabilistic inference by sequential adjustment makes the interesting, empirically testable prediction that the less time and computation a person invests into generating an estimate, the more biased her estimate will be towards the anchor. As illustrated in Fig. 6a, the relative adjustment (see Fig. 5) increases with the number of adjustments. When the number of adjustments is zero, then the relative adjustment is zero and the prediction is the anchor regardless of how far it is away from the correct value. However, as the number of adjustments increases, the relative adjustment increases and the predictions become more informed by the correct value. As the number of adjustments tends to infinity, the average guess generated by anchoring-and-adjustment converges to the expected value of the posterior distribution.

Our analysis of optimal resource-allocation shows that, for a wide range of plausible costs of computation, the resource-rational number of adjustments is much smaller than the number of adjustments required for convergence to the posterior distribution. This might explain why people's estimates of unknown quantities are biased towards their anchor across a wide range of circumstances. Yet, optimal resource allocation also entails that the number of

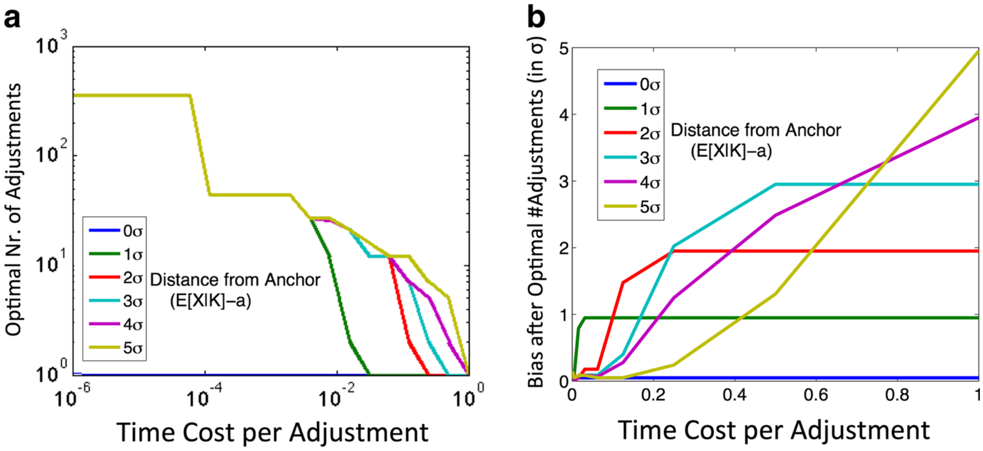

Fig. 4 Optimal number of adjustments (**a**) and the bias after optimal number of adjustments (**b**) as a function of relative time cost and distance from the anchor

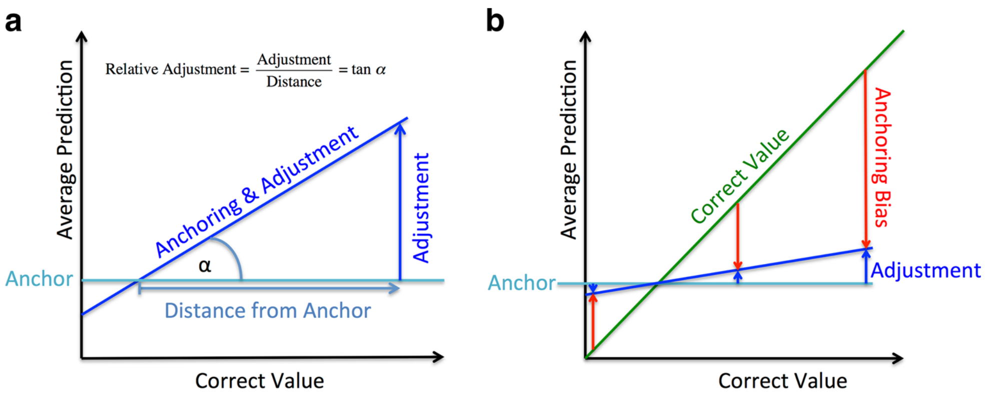

Fig. 5 If the relative adjustment is less than 100%, then the adjustment is less than the distance from the anchor and the prediction is biased (Panel **a**) and the magnitude of the anchoring bias increases with the distance of the correct value from the anchor (Panel **b**)

adjustments increases with the relative cost of error and decreases with the relative cost of time. Hence, our theory predicts that the anchoring bias is smaller when errors are costly and larger when time is costly; Fig. 6b illustrates this prediction.

Although we derived the implications of making rational use of finite cognitive resources for a specific computational mechanism based on sampling, the crucial property of diminishing returns per additional computation is a universal feature of iterative inference mechanisms including approximate Bayesian computation (Sunnåker et al., 2013; Turner & Sederberg, 2012), (stochastic) gradient descent, variational Bayes, predictive coding (Friston & Kiebel, 2009; Friston, 2009), and probabilistic computation in cortical microcircuits (Habenschuss et al., 2013). Therefore, the qualitative predictions shown in Figs. 3–6 are not specific to the abstract computational architecture that we chose to analyze but characterize bounded rationality for a more general class of cognitive architectures. Hence, while we do *not* claim that the brain implements the sampling algorithm we have analyzed, there are many biologically and psychologically plausible mechanisms that share the same characteristics. We will elaborate on this idea in the General Discussion. In the following sections, we assess these and other predictions of our model.

Simulation of anchoring effects

Having derived a resource-rational model of anchoring-and-adjustment we performed computer simulations to test whether this model is sufficient to explain the plethora of anchoring effects reviewed above. To capture our assumption that people make adjustments in discrete steps, we model the size of adjustments using the Poisson distribution $P(\delta) = \text{Poisson}(|\delta|; \mu_{\text{prop}})$. The simulated effects cover a wide range of different phenomena, and our goal is to account for all of these phenomena with a single model.

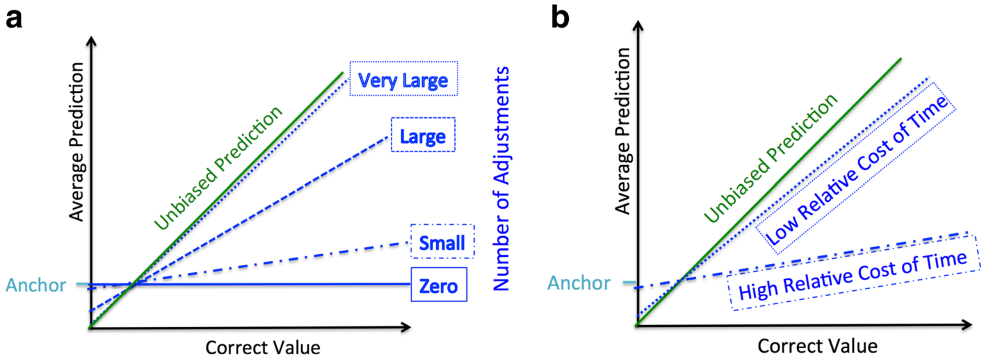

Fig. 6 The number of adjustments increases the relative size of adjustments (*left* panel). As the relative cost of time increases, the number of adjustments decreases and so does the relative size of the adjustment (*right* panel)

Simulation methodology

We simulated the anchoring experiments listed in Table 1 with the resource-rational anchoring-and-adjustment model described above. The participants in each of these experiments were asked to estimate the value of one or more quantities X; for instance (Tversky & Kahneman, 1974) asked their participant to estimate the percentage of African countries in the United Nations. Our model's prediction of people's estimates of a quantity X depends on their probabilistic belief $P(X|K)$ based on their knowledge K, the number of adjustments, the anchor, and the adjustment step-size. Thus, before we could apply our model to simulate anchoring experiments, we had to measure people's probabilistic beliefs $P(X|K)$ about the quantities used on the simulated experiments. Appendix C describes our methodology and reports the estimates with obtained.

To accommodate differences in the order of magnitude of the quantities to be estimated and the effect of incentives for accuracy, we estimated two parameters for each experiment: the expected step-size μ_{prop} of the proposal distribution $P(\delta) = \text{Poisson}(|\delta|; \mu_{\text{prop}})$ and the relative iteration cost γ. These parameters were estimated by the ordinary least-squares method applied to the summary statistics reported in the literature. For experiments comprising multiple conditions using the same questions with different incentives for accuracy we estimated a single step-size parameter that is expected to apply across all conditions and a distinct relative time cost parameter for each incentive condition.

Insufficient adjustment from provided and self-generated anchors

Resource-rational anchoring-and-adjustment provides a theoretical explanation for insufficient adjustment from provided and self-generated anchors in terms of a rational speed-accuracy tradeoff, but how accurately does this describe empirical data? To answer this question, we fit our model to two well-known anchoring experiments: one with provided and one with self-generated anchors.

Provided anchors As an example of adjustment from provided anchors, we chose the study by Jacowitz and Kahneman (1995), because it rigorously quantifies the anchoring bias. Jacowitz and Kahneman (1995) asked their participants two questions about each of several unknown quantities: First they asked whether the quantity is larger or smaller than a certain value–the *provided anchor*. Next they asked the participant to estimate that quantity. For the first half of the participants the anchor was a low value (i.e. the 15th percentile of estimates people make when no anchor is provided), and for the second half of the participants the anchor was a high value (i.e. the 85th percentile). People's estimates were significantly higher when the anchor was high than when it was low. Jacowitz and Kahneman (1995) quantified this effect by the anchoring index (AI), which is the percentage of the distance from the low to the high anchor that is retained in people's estimates:

$$\text{AI} = \frac{\text{Median}(\hat{X}_{\text{high anchor}}) - \text{Median}(\hat{X}_{\text{low anchor}})}{\text{high anchor} - \text{low anchor}} \cdot 100\% \tag{5}$$

Jacowitz and Kahneman (1995) found that the average anchoring index was about 50%. This means that the difference between people's estimates in the high versus the low anchor condition retained about half of the distance between the two anchors.

We determined the uncertainty σ for each of the 15 quantities by the elicitation method described above. Since Jacowitz and Kahneman (1995) measured people's median estimates in the absence of any anchor, we used those values as our estimates of the expected values μ, because their sample and its median estimates were significantly different from ours.

Next, we estimated the adjustment step-size parameter and the relative time cost parameter by minimizing the sum of squared errors between the predicted and the observed anchoring indices. According to the estimated parameters, people performed 29 adjustments with an average step-size of 22.4 units. With these two estimated parameters the model accurately captures the insufficient adjustment from provided anchors reported by Jacowitz and Kahne (1995): The model's adjustments are insufficient (i.e. anchoring index > 0; see Eq. 5) on all questions for which this had been observed empirically but not for the question on which it had not been observed; see Fig. 7. Our model also captured the magnitude of the anchoring bias: the model's average anchoring index of 53.22% was very close to its empirical counterpart of 48.48%. Furthermore, our model also captured for which questions the anchoring bias was high and for which it was low: the correlation between the predicted and the empirical anchoring indices ($r(13) = 0.62$, $p = 0.0135$). The simulated and empirical anchoring effects are shown in Fig. 7.

Self-generated anchors As an example of adjustment from self-generated anchors we chose the studies reported in Epley and Gilovich (2006). In each of these studies participants were asked to estimate one or more unknown quantities such as the boiling point of water on Mount Everest for which many participants readily retrieved a

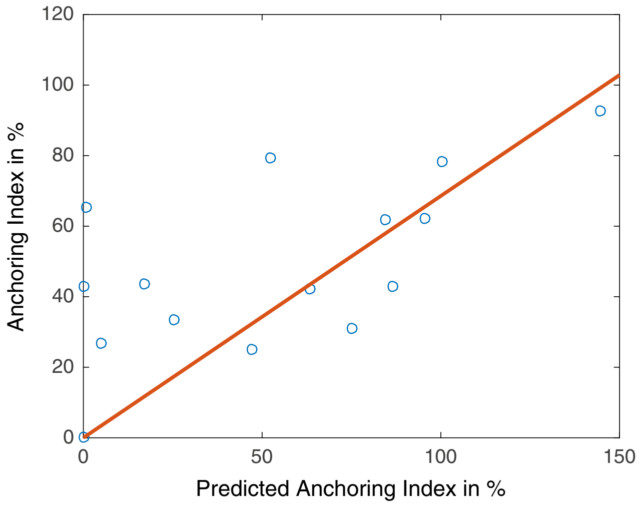

Fig. 7 Simulation of the provided anchor experiment by Jacowitz and Kahneman (1995)

well-known related quantity such as 272 °F (100 °C). Afterwards participants were asked whether they knew and had thought of each intended anchor while answering the corresponding question. For each question, Epley and Gilovich (2006) computed the mean estimate of those participants who had thought of the intended anchor while answering it. We combined the data from all self-generated anchor questions without additional experimental manipulations for which Epley and Gilovich (2006) reported people's mean estimate, i.e. the first five question from Study 1a, the first five questions from Study 1b, and the control conditions of Study 2b (2 questions) and the first seven questions from Study 2c.[2] We determined the means and uncertainties of the model's beliefs about all quantities used in Epley and Gilovich's studies by the elicitation method described above. The anchors were set to the intended self-generated anchors reported by Epley and Gilovich (2006). We estimated the model's time cost and adjustment step-size parameters by fitting the relative adjustments reported for these studies using the ordinary least-squares method.

The estimated parameters suggest that people performed 8 adjustments with an average step-size of 10.06 units. With these parameters the model adjusts its initial estimate by 80.62% of the distance to the correct value; this is very close to the 80.95% relative adjustment that Epley and Gilovich (2006) observed on average across the simulated studies. Our model captures that for the majority of quantities (13 out of 19) people's adjustments were insufficient. It also captures for which questions people adjust more and for which questions they adjust less from their uncertainties and anchors: as shown in Fig. 8 our model's predictions of the relative adjustments were significantly correlated with the relative adjustments that Epley and Gilovich (2006) observed across different questions ($r(17) = 0.61$, $p = 0.0056$). Comparing the parameter estimates between the experiments with provided versus self-generated anchors suggests that people adjusted less when they had generated the anchor themselves. This makes sense because self-generated anchors are typically much closer to the correct value than provided anchors.

Effect of cognitive load

In an experiment with self-generated anchors (Epley & Gilovich, 2006) found that people adjust their estimate less when required to simultaneously memorize an eight-letter string. To investigate whether resource-rational anchoring-and-adjustment can capture this effect, we fit our model simultaneously to participants' relative adjustment with versus without cognitive load. Concretely, we estimated a common step-size parameter and separate time cost parameters for each condition by the least squares method. We included all items for which Epley and Gilovich (2006) reported people's estimates. The resulting parameter estimates captured the effect of cognitive load: when people were cognitively busy, the estimated cost per adjustment was 4.58% of the error cost, but when people were not cognitively busy then it was only 0.003% of the error cost. The estimated average step-size per adjustment was $\mu = 11.69$. According to these parameters participants performed only 14 adjustments when they were under cognitive load but 60 adjustments when they are not. With these parameters

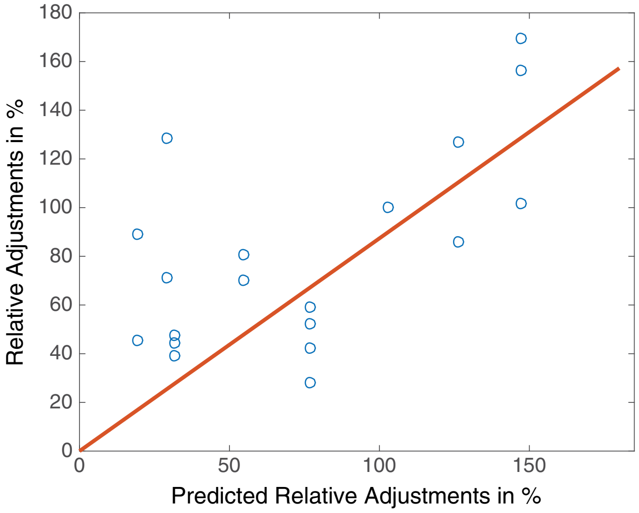

Fig. 8 Simulation of self-generated anchors experiment by Epley and Gilovich (2006)

[2] The quantities were the year in which Washington was elected president, the boiling point on Mt. Everest, the freezing point of vodka, the lowest body temperature, the highest body temperature, and the duration of pregnancy in elephants. Some of these quantities were used in multiple studies.

our model captures the effect of cognitive load on relative adjustment: cognitive load reduced the simulated adjustments by 18.61% (83.45% under load and 102.06% without load). These simulated effects are close to their empirical counterparts: people adjusted their estimate 72.2% when under load and 101.4% without cognitive load (Epley & Gilovich, 2006). Furthermore, the model accurately captured for which questions the effect of cognitive load was high and for which it was low; see Fig. 9. Concretely, our model explained 93.03% of the variance in the effect of cognitive load on relative adjustments ($r(5) = 0.9645$, $p < 0.001$).

The anchoring bias increases with anchor extremity

Next we simulated the anchoring experiment by Russo and Schoemaker (1989). In this experiment business students were first asked about the last three digits of their telephone number. Upon hearing the number the experimenter announced he would add 400 to this number (providing an anchor) and proceeded to ask the participant whether the year in which Attila the Hun was defeated in Europe was smaller or larger than that sum. When the participant indicated her judgment, she was prompted to estimate the year in which Attila had actually been defeated. Russo and Schoemaker (1989) then compared the mean estimate between participants whose anchor had been 500 ± 100, 700 ± 100, $\cdots$, 1300 ± 100. They found that their participants' mean estimates increased linearly with the provided anchor even though the correct value was A.D. 451.

To simulate this experiment, we determined the values of μ and σ by the elicitation method described above. Since the variability of people's estimates and confidence intervals was very high, we increased the sample size of this one experiment to 200. We set the model parameters to the values estimated from the provided anchor experiments by Jacowitz and Kahneman (1995) (see above). As Fig. 10 shows, our model correctly predicted that people's estimates increase linearly with the provided anchor (Russo & Schoemaker, 1989). To determine whether the quantitative differences between the model predictions and the data reported by Russo and Schoemaker (1989) were due to differences between business students in 1989 and people working on Mechanical Turk in 2014, we ran an online replication of their experiment on Mechanical Turk with 300 participants. There appeared to be no significant difference between the estimates of the two populations. However, people's estimates were highly variable. Consequently, the error bars on the mean estimates are very large.

Taking into account the high variance in people's judgments, our simulation results are largely consistent with the empirical data. In particular, both Russo and Shoemaker's

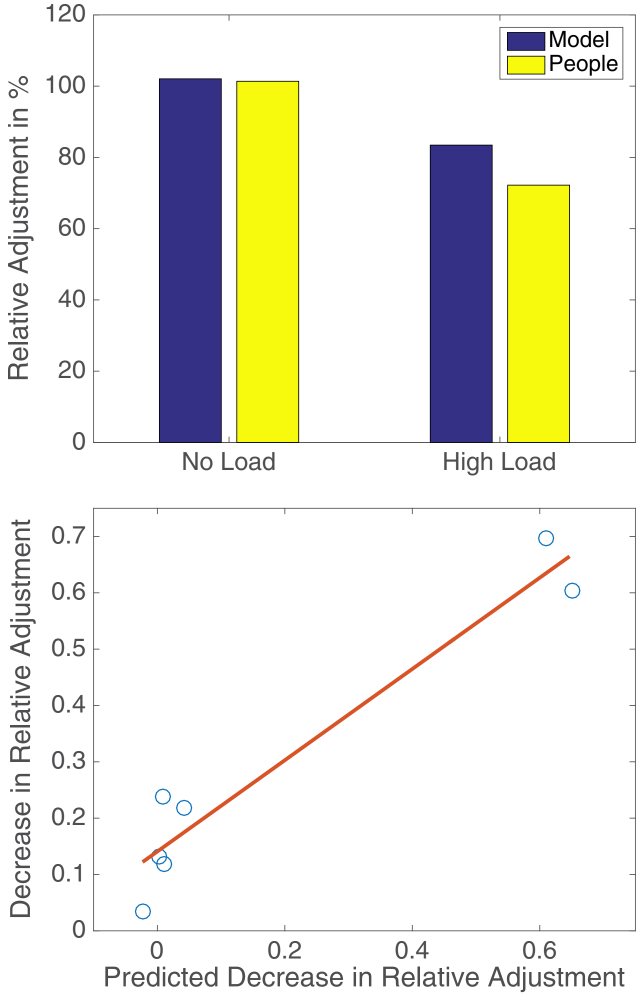

Fig. 9 Simulated versus observed effect of cognitive load on the size of people's adjustments

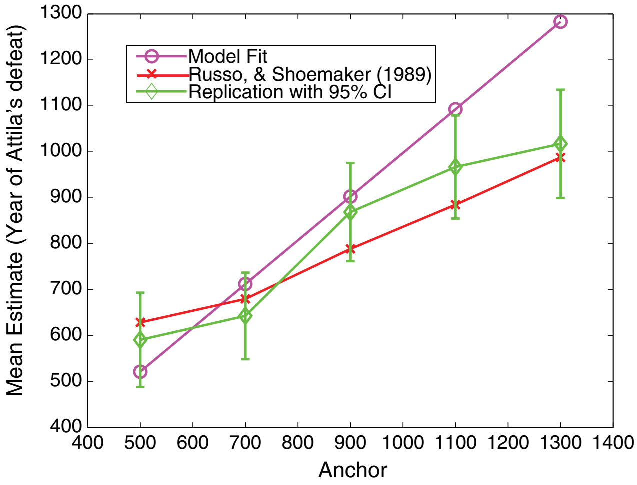

Fig. 10 Simulated effect of the anchor on people's estimates of the year of Atilla's defeat and empirical data from Russo & Shoemaker (1989)

data and our replication confirm our model's qualitative prediction that the magnitude of the anchoring bias increases linearly with the anchor, although our model's prediction for the highest anchor was more extreme than the average judgment.

The effects of uncertainty and knowledge

Several experiments have found that the anchoring bias is larger the more uncertain people are about the quantity to be estimated (Wilson et al. 1996; Jacowitz & Kahneman 1995). To assess whether and how well our theory can explain this effect, we re-analyzed our simulation of the experiment by Jacowitz and Kahneman (1995) reported above. Concretely, we computed the correlation between the uncertainties σ of the modeled beliefs about the 15 quantities and the predicted anchoring indices. We found that resource-rational anchoring-and-adjustment predicted that adjustments decrease with uncertainty. Concretely, the anchoring index that our model predicted for each quantity X was significantly correlated with the assumed uncertainty (standard deviation σ) about it (Spearman's $\rho = 0.5857$, $p = 0.0243$). This is a direct consequence of our model's probabilistic acceptance or rejection of proposed adjustments on a flat (high uncertainty) versus sloped (low uncertainty) belief distribution $P(X|K) = \mathcal{N}(\mu, \sigma)$. Our model thereby explains the negative correlation ($r(13) = -0.68$) that Jacowitz and Kahneman (1995) observed between confidence ratings and anchoring indices.

Uncertainty reflects the lack of relevant knowledge. Thus people who are knowledgeable about a quantity should be less uncertain and consequently less susceptible to anchoring. Wilson et al. (1996) conducted an anchoring experiment in which people first compared the number of countries in the United Nations (UN) to an anchor, then estimated how many countries there are in the UN, and finally rated how much they know about this quantity. They found that people who perceived themselves as more knowledgeable were resistant to the anchoring bias whereas people who perceived themselves as less knowledgeable were susceptible to it. Here, we asked whether our model can explain this effect by smaller adjustments due to higher uncertainty. To answer this question, we recruited 60 participants on Mechanical Turk, asked them how much they knew about the number of nations in the UN on a scale from 0 ("nothing") to 9 ("everything") and elicited their beliefs by the method described in Appendix C. We then partitioned our participants into a more knowledgeable and a less knowledgeable group by a median split as in Wilson et al. (1996). We model the beliefs elicited from the two groups by two separate normal distributions (Appendix C).

We found that the high-knowledge participants were less uncertain than the low-knowledgeable participants ($\sigma_{\text{high}} = 35.1$ vs. $\sigma_{\text{low}} = 45.18$). Furthermore, their median estimate was much closer to the true value of 193 ($\mu_{\text{high}} = 185$ vs. $\mu_{\text{low}} = 46.25$). We fit the relative adjustments from the anchor provided in Wilson et al.'s experiment (1930) by the least-squares method as above. With the estimated parameters (17 adjustments, step-size 488.2) the model's predictions captured the effect of knowledge: For the low-knowledge group the model predicted that providing the high anchor would raise their average estimate from 45.18 to 252.1. By contrast, for the high-knowledgeable group our model predicted that providing a high anchor would fail to increase people's estimates (185 without anchor, 163 with high anchor).

Differential effects of accuracy motivation

People tend to invest more mental effort when they are motivated to be accurate. To motivate participants to be accurate some experiments employ financial incentives for accuracy, while others warn their participants about potential errors that should be avoided (forewarnings). Consistent with the effect of motivation, resource-rational anchoring-and-adjustment predicts that the number of adjustments increases with the relative cost of error. Yet, financial incentives for accuracy reduce the anchoring bias only in some circumstances but not in others: First, the effect of incentives appeared to be absent when anchors were provided but present when they were self-generated (Tversky and Kahneman, 1974; Epley & Gilovich, 2005). Second, the effect of incentives was found to be larger when people were told rather than asked whether the correct value is smaller or larger than the anchor (Simmons et al., 2010). Here, we explore whether and how these interaction effects can be reconciled with resource-rational anchoring-and-adjustment.

Smaller incentive effects for provided anchors than for self-generated anchors Epley and Gilovich (2005) found that financial incentives and forewarnings decreased the anchoring bias when the anchor was self-generated but not when it was provided by the experimenter. From this finding Epley and Gilovich (2005) concluded that people use anchoring-and-adjustment only when the anchor is self-generated but not when it is provided. By contrast, Simmons et al. (2010) suggested that this difference may be mediated by people's uncertainty about whether the correct answer is larger or smaller than the anchor. They found that people are often uncertain in which direction they should adjust in questions used in experiments with provided anchors; so this may be why incentives for accuracy failed to reduce the anchoring bias in those experiments. Here we show that resource-rational anchoring-and-adjustment can capture the differential effectiveness of financial incentives in experiments with provided versus self-generated anchors. First, we show through simulation that given the amount of uncertainty that people have about the quantities to be estimated our model predicts a larger effect of accuracy motivation for the self-generated anchor experiments by Epley and Gilovich (2005) than for the provided anchor experiments by Tversky and Kahneman (1974) and Epley and Gilovich (2005).

First, we analyze people's beliefs about the quantities used in experiments with provided versus self-generated anchors with respect to their uncertainty. We estimated the mean μ and standard deviation σ of people's beliefs about each quantity X by the elicitation method described above. Because the quantities' values differ by several orders of magnitude, it would be misleading to compare the standard deviations directly. For example, for the population of Chicago (about 2,700,000 people) a standard deviation of 1,000 would express near-certainty, whereas for the percentage of countries in the UN the same standard deviation would express complete ignorance. To overcome this problem, the standard deviation has to be evaluated relative to the mean. We therefore compare uncertainties in terms of the signal-to-noise ratio (SNR). We estimated the SNR by the median of the signal-to-noise ratios of our participants' beliefs ($\mathrm{SNR}_s = \mu_s^2/\sigma_s^2$). We found that people tended to be much more certain about the quantities Epley and Gilovich (2005) used in their self-generated anchors experiments (median SNR: 21.03) than about those for which they provided anchors (median SNR: 4.58). A Mann-Whitney U-test confirmed that the SNR was significantly higher for self-generated anchoring questions than for questions with provided anchors ($U(18) = 74.0$, $p = 0.0341$).

Given that people were more uncertain about the quantities used in the experiments with provided anchors, we investigated how this difference in uncertainty affects the effect of financial incentives on the anchoring bias predicted by our resource-rational model. To do so, we simulated Study 1 from Epley and Gilovich (2005), in which they compared the effects of financial incentives between questions with self-generated versus provided anchors, and the provided anchors experiment by Tversky and Kahneman (1974). To assess whether our model can explain why the effect of motivation differs between questions with provided versus self-generated anchors, we evaluated the effects of motivation as follows: First, we fit our model to the data from the condition with self-generated anchors. Second, we use the estimated numbers of adjustments to simulate responses in the condition with provided anchors. Third, for each question, we measured the effect of motivation by the relative adjustment with incentives minus the relative adjustment without incentives. Fourth, we averaged the effects of motivation separately for all questions with self-generated versus provided anchors and compared the results.

We fit the relative adjustments on the questions with self-generated anchors with one step-size parameter and two relative time-cost parameters: The estimated step-size was 17.97. The estimated number of adjustments was 5 for the condition without incentives and 9 for the condition with incentives. According to these parameters, motivation increased the relative adjustment from self-generated anchors by 12.74% from 65.62% to 78.35%. This is consistent with the significant effect of 33.01% more adjustment that Epley and Gilovich (2005) observed for questions with self-generated anchors. For the condition with provided anchors Epley and Gilovich (2005) used four questions from the experiment by Jacowitz and Kahneman (1995) simulated above and the same incentives as in the questions with self-generated anchors. We therefore simulated people's responses to questions with provided anchors using the step-size estimated from the data by Jacowitz and Kahneman (1995) and the number of adjustments estimated from questions with self-generated anchors. Our simulation correctly predicted that incentives for accuracy fail to increase adjustment from provided anchors. Concretely, our simulation predicted 44.09% adjustment with incentives and 44.48% without. Thus, as illustrated in Fig. 11, our model captures that financial incentives increased adjustment from self-generated anchors but not from provided anchors. According to our model, this difference is just an artifact of the confound that people know more about the quantities used in experiments with self-generated anchors than about the quantities used in experiments with provided anchors.

Finally, we simulated Study 2 from Epley and Gilovich (2005) in which they compared the effect of warning participants about the anchoring bias between questions with provided versus self-generated anchors. This study had 2 (self-generated anchors vs. provided anchors) × 2 (forewarnings vs. no forewarnings) conditions. Epley and Gilovich (2005) found that in the conditions with self-generated anchors forewarnings increased adjustment, but in the conditions with provided anchors they did not. As before, we set the model's beliefs about the quantities used in this experiment using the elicitation method described above. We fit our model to the relative adjustments in the conditions with self-generated anchors. Concretely, we used the least-squares method to fit one step-size parameter and two time cost parameters: one for the condition with forewarnings and one for the condition without forewarnings. With these parameters, we simulated people's estimates in the conditions with self-generated anchors (to which the parameters were fit) and predicted the responses in the provided anchor conditions that we had *not* used for parameter estimation.

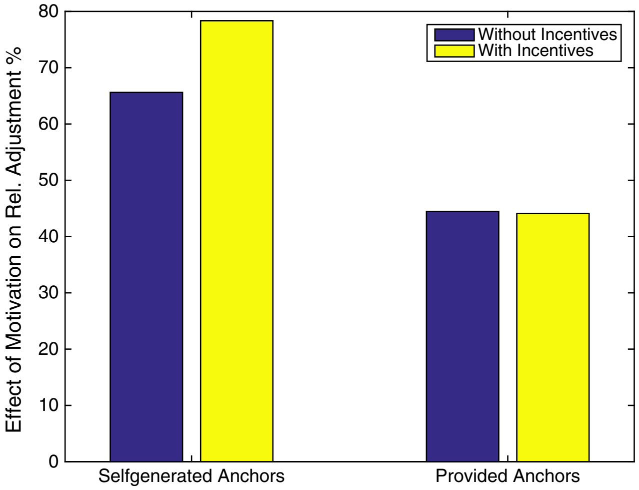

Fig. 11 Simulation of Study 1 from Epley and Gilovich (2005): Predicted effects of financial incentives on the adjustment from provided versus self-generated anchors

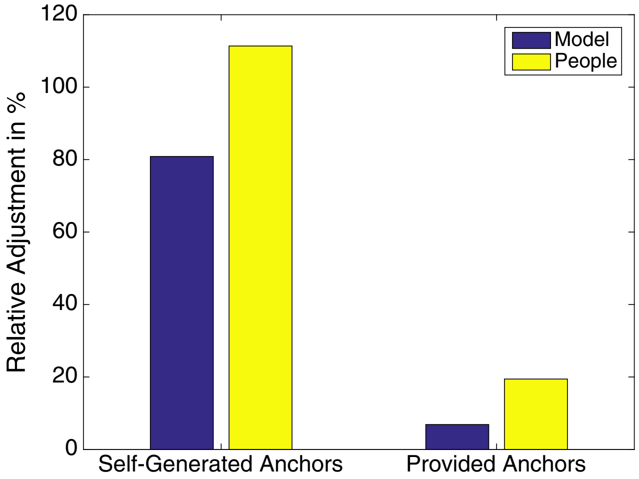

Fig. 12 Simulation of Study 2 from Epley and Gilovich (2005): Predicted effects of forewarnings for questions from experiments with provided versus self-generated anchors

According to the estimated parameters, forewarnings increased the number of adjustments from 8 to 28. We therefore simulated the responses in both conditions with forewarnings (provided and self-generated anchor questions) with 8 adjustments and all responses in the two conditions without forewarnings (provided and self-generated anchor questions) with 28 adjustments. For the questions with self-generated anchors, forewarnings increased the simulated adjustments by 30% from insufficient 81% to overshooting 111% of the total distance from the anchor to the correct value.[3] By contrast, for questions with provided anchors forewarnings increased the simulated adjustments by only 12.5% from 6.9% to 19.4%. Thus, assuming that forewarnings increase the number of adjustments from provided anchors by the same number as they increase adjustments from self-generated anchors our model predicts that their effect on people's estimates would be less than one third of the effect for self-generated anchors; see Fig. 12. According to our model, the reason is that people's uncertainty about the quantities for which anchors were provided is so high that the effect of additional adjustments is much smaller than in the questions for which people can readily generate their own anchors. Our results are consistent with the interpretation that the absence of a statistically significant effect of forewarnings on the bias towards the provided anchors in the small sample of Epley and Gilovich (2005) does not imply that the number of adjustments did not increase. Therefore adjustment from provided anchors cannot be ruled out.

Direction uncertainty masks the effect of incentives

Simmons et al. (2010) found that accuracy motivation decreases anchoring if people are confident about whether the quantity is larger or smaller than the anchor but not when they are very uncertain. Simmons et al. (2010) showed that even when the anchor is provided, incentives for accuracy can reduce the anchoring bias provided that people are confident about the correct direction of adjustment. Concretely, Simmons et al.'s second study unmasked the effect of incentives on adjustment from provided anchors by telling instead of asking their participants whether the true value is larger or smaller than the anchor. Similarly, in their third study, Simmons et al. (2010) found that the effect of incentives is larger when the provided anchor is implausibly extreme than when it is plausible. Here we report simulations of both of these effects.

First, we show that our model can capture that the effect of incentives increases when people are told the correct direction of adjustment. Simmons et al.'s second study measured the effect of accuracy motivation on the anchoring index as a function of whether people were asked or told if the correct value is larger or smaller than the anchor. We modeled the effect of being told that the quantity X is smaller or larger than the anchor a by Bayesian updating of the model's belief about X from $P(X|K)$ to $P(X|K, X < a)$ and $P(X|K, X > a)$ respectively. The original beliefs $P(X|K)$ were determined by the elicitation method described in Appendix C. We fit the model simultaneously to all anchoring indices by ordinary least squares to estimate one step-size parameter and one number of adjustments for each incentive condition. According to the estimated parameters, incentives increased the number of adjustments from 5 to 1000 and the average adjustment step-size was

[3] Overshooting is possible, because the expected value of the estimated belief $P(X|K) = \mathcal{N}(\mu, \sigma)$ can be farther away from the anchor than the correct value.

11.6 units. For both incentive conditions, our model captured the variability of adjustments across trials: For trials with incentives for accuracy the correlation between simulated and measured anchoring indices was $r(18) = 0.77$ ($p = 0.0001$), and for trials without incentives this correlation was $r(18) = 0.61$ ($p = 0.004$). Our model also captured the overall reduction of anchoring with incentives for accuracy observed by Simmons et al. (2010), although the predicted 42% reduction of anchoring with incentives for accuracy was quantitatively larger than the empirical effect of 8%. Most importantly, our model predicted the effects of direction uncertainty on adjustment and its interaction with accuracy motivation: First, our model predicted that adjustments are larger if people are told whether the correct value is larger or smaller than the anchor. The predicted 13.7% reduction in the anchoring index was close to the empirically observed reduction by 18.8%. Second, our model predicted that the effect of accuracy motivation will be 6.3% larger when people are told the direction of adjustment. The predicted effect of direction uncertainty is smaller than the 21% increase reported by Simmons et al. (2010) but qualitatively consistent. Therefore, our model can explain why telling people whether the correct value is larger or smaller than the anchor increases the effect of accuracy motivation. According to our model, financial incentives increase the number adjustments in both cases, but knowing the correct direction makes adjustment more effective by eliminating adjustments in the wrong direction.

Second, we simulated Study 3b of Simmons et al. (2010) in which they showed that financial incentives increase adjustments away from implausible anchors. Concretely, this study compared the effect of accuracy motivation on adjustments between plausible versus implausible provided anchors. As before, we determined the model's beliefs by the procedure described above and estimated the number of adjustments with and without incentives (781 and 188) and the adjustment step-size (0.01) by fitting the reported relative adjustments by ordinary-least squares.[4] With this single set of parameters we simulated adjustments from plausible versus implausible provided anchors. The predicted adjustments captured a statistically significant proportion of the effects of anchor type, motivation, and quantity on the size of people's adjustments: $\rho(22) = 0.72$, $p < 0.0001$. Most importantly, our simulations predicted no statistically significant effect of accuracy motivation on absolute adjustment (mean effect: 0.76 units; 95% CI: [−0.42; 1.94]) when the anchor was plausible but a substantially larger and statistically significant effect when the anchor was implausible (17.8 units; 95% CI: [9.76; 25.91]); see Fig. 13. This prediction results from the fact that large adjustments away from plausible anchors will often be rejected because they decrease the estimate's plausibility and small adjustments in the wrong direction are almost as likely to be accepted as adjustment in the correction direction because values on either side of the plausible anchor are almost equally plausible if the distribution is symmetric around its mode. Thus the expected change per adjustment is rather small.

In conclusion, resource-rational anchoring-and-adjustment can explain why motivating participants to be accurate reduces the anchoring bias in some circumstances but not in others. In a nutshell, our model predicts that incentives for accuracy have little effect when adjustments in either direction hardly change the estimate's plausibility. The simulations reported above demonstrate that this principle is sufficient to explain the differential effect of accuracy motivation on adjustments from provided versus self-generated anchors. Therefore, a single process – resource-rational anchoring-and-adjustment – may be sufficient to explain anchoring on provided and self-generated anchors.

Summary

Our resource-rational analysis of numerical estimation showed that under-adjusting an initial estimate can be a rational use of computational resources. The resulting model can explain ten different anchoring phenomena: insufficient adjustments from both provided and self-generated anchors, the effects of cognitive load, anchor extremity, uncertainty, and knowledge, as well as the differential effects of forewarnings and financial incentives depending on anchor

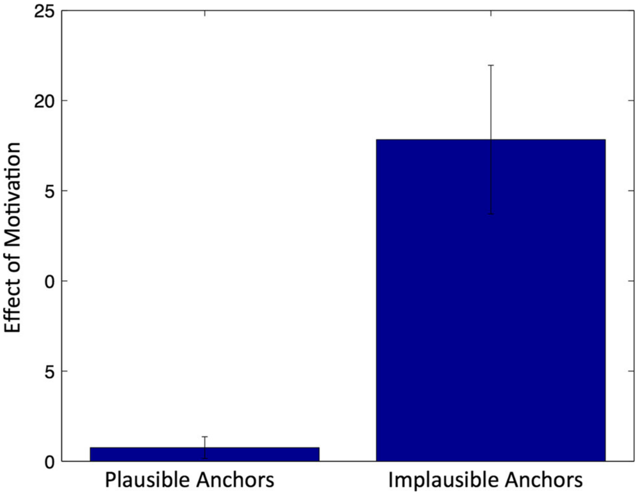

Fig. 13 Simulation of Experiment 3 from Simmons et al. (2010): Predicted effect of accuracy motivation on adjustments from plausible versus implausible provided anchors

[4] The reason that the estimated step-size is so small appears to be that all quantities and distances in this experiment are small compared to those in other experiments such as Study 2 by the same authors. The increase in the number of adjustments appears to compensate for the reduced step-size.

type (provided vs. self-generated), anchor plausibility, and being asked versus being told whether the quantity is smaller or larger than the anchor (see Table 1). None of the previous models (Epley and Gilovich, 2006; Simmons et al., 2010) was precise enough to make quantitative predictions about any of these phenomena let alone precisely predict all of them simultaneously. The close match between our simulation results and human behavior suggests that resource-rational anchoring-and-adjustment provides a unifying explanation for a wide range of disparate and apparently incompatible phenomena in the anchoring literature. Our model was able to reconcile these effects by capturing how the effect of adjustment depends on the location and shape of the posterior distribution describing the participants' belief about the quantity to be estimated. For instance, our model reconciles the apparent ineffectiveness of financial incentives at reducing the bias towards provided anchors (Tversky & Kahneman, 1974) with their apparent effectiveness at reducing bias when the anchor is self-generated (Epley & Gilovich, 2005). To resolve this apparent contradiction, we did not have to postulate additional processes that operate only when the anchor is provided–unlike Epley and Gilovich (2006). Instead, our computational model directly predicted this difference from people's higher uncertainty about the quantities used in experiments with provided anchors, because when the uncertainty is high then adjustments in the wrong direction are more likely to be accepted. Our model thereby provides a more parsimonious explanation of these effects than the proposal by Epley and Gilovich (2006). While Simmons et al. (2010) offered a conceptual explanation along similar lines, our model predicted the exact sizes of these effects a priori.

The parameter estimates we obtained differed significantly across the simulated phenomena. This is partly due differences in the incentives and other experimental manipulations. Additional reasons for the variability in the parameter estimates are somewhat arbitrary differences in the resolution of the hypothesis spaces across different quantities and the interdependence between the average change per adjustment and the number of adjustments: the same amount of adjustment can be explained either by a small number of large steps or a large number of small steps. For some experiments maximum likelihood estimation chose the former interpretation and for others it chose the latter. But because a larger step size can compensate for a smaller number of adjustments, it is quite possible that the model could have explained all of the findings with a very similar step size and number of adjustment parameters if we knew the structure and resolution of people's hypothesis spaces for the quantities used in each experiment. Although the model's parameters were unknown and had to be estimated to make quantitative predictions, all of the qualitative phenomena we simulated logically follow from the structure of the model itself. In this sense, our model did not just capture the simulated phenomena but predicted them. Most importantly, our theory reconciles the apparently irrational effects of potentially irrelevant numbers with people's impressive capacity to efficiently handle a large number of complex problems full of uncertainty in a short amount of time.

General discussion

Anchoring and adjustment is one of the classic heuristics reported by Tversky and Kahneman (1974) and it seems hard to reconcile with rational behavior. In this article, we have argued that this heuristic can be understood as a signature of resource-rational information processing rather than a sign of human irrationality. We have supported this conclusion by a resource-rational analysis of numerical estimation and simulations of ten anchoring phenomena with a resource-rational process model. We showed that anchoring-and-adjustment can be interpreted as a Markov chain Monte Carlo algorithm–a rational approximation to rational inference. We found that across many problems the optimal speed-accuracy tradeoff of this algorithm entails performing so few adjustments that the resulting estimate is biased towards the anchor. Our simulations demonstrated that resource-rational anchoring-and-adjustment, which adaptively chooses the number of adjustments to maximize performance net the cost of computation, provides a unifying explanation for ten different anchoring phenomena (see Table 1).

Although we explored the implications of limited time and finite cognitive resources assuming an abstract computational architecture based on sampling, we do *not* claim that the brain implements the sampling algorithm we analyzed above. Instead, our goal was to illustrate general properties of resource-rational information processing. Many other iterative inference mechanisms also have the property of diminishing returns for additional computation that our analysis is based on. Hence, the qualitative predictions shown in Figs. 3–6 characterize bounded rationality for a more general class of cognitive architectures. Importantly, this class includes biologically plausible neural network implementations of Bayesian inference (Habenschuss et al., 2013; Friston and Kiebel, 2009; Friston, 2009) and mechanisms that implement the general principles of our model in a more psychologically plausible fashion. For instance, while our model's assumption that people can evaluate the exact likelihood of the observed data under each sampled hypothesis is questionable, our analysis also applies to sampling methods that approximate the likelihood through simulation (Turner & Sederberg, 2012; Sunnåker et al., 2013). Likewise, while we do not propose a neural implementation of probabilistic inference, our analysis also applies to

Markov chain Monte Carlo algorithms implemented in cortical microcircuits (Habenschuss et al., 2013), stochastic gradient descent, and the predictive coding implementation of variational inference postulated by the free-energy principle (Friston & Kiebel, 2009; Friston, 2009). Therefore, our results support the adaptive allocation of finite computational resources and the resource-rationality of bias regardless of the specific cognitive mechanism that people use to draw inferences.

In the remainder of this paper we will discuss the implications of our results for general theoretical questions. We start by discussing how our model is related to previous theories of anchoring and how they can be integrated into our resource-rational framework. We then turn to two questions about rationality: First, we discuss existing evidence for the hypothesis that anchors are chosen resource-rationally and how it can be tested in future experiments. Second, we argue that resource-rationality, the general theory we have applied to explain the anchoring bias, provides a more adequate normative framework for cognitive strategies than classical notions of rationality. We close with directions for future research.

Relation to previous theories of anchoring and adjustment

Previous models of anchoring-and-adjustment (Epley & Gilovich, 2006; Simmons et al., 2010) assumed that adjustment terminates when the plausibility of the current estimate exceeds a threshold. From an information processing perspective, the limitation of models postulating that adjustment stops when plausibility exceeds a threshold is that there is no single threshold that works well across all estimation problems. Depending on the level of uncertainty successful estimation requires different thresholds. A threshold that is appropriate for low uncertainty will result in never-ending adjustment in a problem with high uncertainty. Conversely, a threshold that is appropriate for a problem with high uncertainty would be too liberal when the uncertainty is low. In addition, Simmons et al. (2010) postulate that people reason about the direction of their adjustment whereas resource-rational anchoring-and-adjustment does not. It would be interesting to see whether an extension of our model that incorporates directional information would perform better in numerical estimation and better predict human behavior. We will return to this idea when we discuss directions for future research.

According to the selective-accessibility theory of anchoring (Strack and Mussweiler, 1997), comparing an unknown quantity to the provided anchor increases the accessibility of anchor-consistent knowledge and the heightened availability of anchor-consistent information biases people's estimates. There is no quantitative mathematical model of selective accessibility that could be tested against our resource-rational anchoring-and-adjustment model using the data we have collected. The evidence that some anchoring biases result from selective accessibility (Strack & Mussweiler, 1997) does not undermine our analysis, because the existence of selective accessibility would not rule out the existence of anchoring-and-adjustment and vice versa. In fact, from the perspective of resource-rational probabilistic inference a mechanism similar to selective accessibility is likely to coexist with anchoring-and-adjustment. Concretely, we have formalized the problem of numerical estimation of some quantity X as minimizing the expected error cost of the estimate $\hat{x}$ with respect to the posterior distribution $P(X|K)$ where K is the entirety of the person's relevant knowledge. This problem can be decomposed into two sub-problems: conditioning on relevant knowledge to evaluate (relative) plausibility and searching for an estimate with high plausibility. It appears unlikely that the mind can solve the first problem by simultaneously retrieving and instantly incorporating each and every piece of knowledge relevant to estimating X. Instead, the mind might have to sequentially recall and incorporate pieces $K^{(1)}, K^{(2)}, K^{(3)}, \cdots$ of its knowledge to refine $P(X)$ to $P(X|K^{(1)})$ to $P(X|K^{(1)}, K^{(2)})$ to $P(X|K^{(1)}, K^{(2)}, K^{(3)})$, and so forth. This process could be modeled as bounded using a sequential Monte Carlo algorithm (Doucet et al., 2001) and bounded conditioning (Horvitz et al., 1989). Furthermore, it would be wasteful not to consider the knowledge that has been retrieved to answer the comparison question in the estimation task and impossible to retrieve all of the remaining knowledge. Selective accessibility may therefore result from the first process. Yet, regardless of how the first problem is solved, the mind still needs to search for an estimate $\hat{x}$ with high posterior probability, and this search process might be implemented by something like anchoring-and-adjustment. Furthermore, the knowledge retrieved in the first step might also guide the generation of an anchor. Importantly, both processes are required to generate an estimate. Therefore, we agree with Simmons et al. (2010) that selective accessibility and anchoring-and-adjustment might coexist and both of them might contribute to the anchoring bias.

Like the model by Simmons et al. (2010), our theory deviates from Epley and Gilovich (2005) by suggesting that anchoring and adjustment is a unifying mechanisms for the anchoring biases observed for self-generated as well as provided anchors. Our simulations show that this assertion is compatible with the results reviewed by Epley and Gilovich (2006) because the effect of financial incentives declines with the uncertainty about the quantity to be estimated. This explanation is similar to the argument by Simmons et al. (2010), but our formal model does not need to assume that people reason about the direction of their adjustments.

Our model is consistent with the recently proposed anchor integration model (Turner and Schley, 2016). Both models describe the effect of the anchor in terms of Bayesian inference, but while the anchor integration model is agnostic about the mechanism by which the anchor affects people's judgments and whether or not this is rational, we have developed a rational process model.

In summary, our resource-rational analysis of estimation problems sheds new light on classic notions of anchoring-and-adjustment (Tversky & Kahneman, 1974; Epley & Gilovich, 2006), explaining why they work and why people use them. Furthermore, our framework is sufficiently general to incorporate and evaluate the additional mechanisms proposed by Simmons et al. (2010) and Strack and Mussweiler (1997) and many others. Exploring these extensions is an interesting direction for future work.

Are anchors chosen rationally?

Anchoring-and-adjustment has two components: generating an anchor and adjusting from it. Our simulations supported the conclusion that adjustment is resource-rational. Thus, a natural next question is whether anchors are also generated resource-rationally.

Self-generated anchors are usually close to the correct value, but provided anchors can be far off. For instance, it appears irrational that people can be anchored on their social security number when they estimate how much they would be willing to pay for a commodity (Ariely et al., 2003). Yet, the strategy failing people in this specific instance may nevertheless be resource-rational overall for at least four reasons: First, it is sensible to assume that the experimenter is reasonable and cooperative. Therefore her utterances should follow the Gricean maxims. Specifically, according to Grice's maxim of relation the stated anchor should be relevant (Zhang & Schwarz, 2013). Furthermore, as a rational information-seeking agent the experimenter should ask the question whose answer will be most informative. The most informative anchor to compare the true value to would be at the center of the experimenter's belief distribution. This too suggests that it is reasonable to treat the provided anchor as a starting point. A weaker version of this argument might apply even to the experiment in which Tversky and Kahneman (1974) asked participants to compare the number of African countries in the UN to a randomly generated number: It seems reasonably for participants to assume that the experimenter would not be asking them whether the number of African countries in the UN is larger or smaller than the number on the wheel of fortune if the answer was obvious to him. Hence, assuming the logic of conversation, the fact that the experimenter did ask would suggest that the number was within the range of values he considered plausible. Under these assumptions, the question constitutes an informative endorsement of the anchor regardless of how it was generated. This makes it reasonable to use that value as a starting point.

Second, subsequent thoughts and questions are usually related. So it is reasonable to use the answer to a preceding question as the starting point for next thought. This holds for sequences of arithmetic operations such as $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ for which people anchor on their intermediate results when they are forced to respond early (Tversky & Kahneman, 1974) and in many other cases too. Third, when the provided anchor is the only number available in working memory, then using it may be faster and require less effort than generating a new one. This assumption is consistent with evidence for spreading-activation theories of semantic processing and memory retrieval (Collins & Loftus, 1975; Neely, 1977; Anderson, 1983). For instance, when primed with one word people are faster to recognize words that are associated with the prime than words that are not (Neely, 1977; Collins & Loftus, 1975). The spreading of activation to associated mental representations appears to be fast and automatic (Neely, 1977) and inhibiting it would be slow and effortful (Diamond, 2013). Furthermore, according to spreading-activation theories of memory recall (Anderson, 1983; Neely, 1977) and rational process models of memory search (Bourgin et al., 2014; Abbott et al., 2015), the generation of a new anchor from memory might be subject to the same limitations as the adjustment process itself. Last but not least, one's beliefs may be wrong and the anchor may be more accurate. This was the case in Russo and Shoemaker's experiment: People overestimated the year in which Attila the Hun was defeated in Europe so much that the anchor was usually closer to the correct value (A.D. 451) than the mean of their unbiased estimates (A.D. 953.5). For these reasons, the observation that people anchor on irrelevant values provided in psychological experiments does not imply that anchors are selected irrationally. Anchor selection could be well adapted to the real-world. Consequently, anchoring biases in everyday reasoning would be much more benign than those observed in the laboratory. This is probably true, because most anchoring experiments violate people's expectation that the experimenter will provide relevant information, provide negligible incentives for accuracy, and ask people to estimate quantities about which they know very little.

There also is empirical evidence suggesting that people do not always use the provided value as their anchor. For instance, in the experiment by Strack and Mussweiler (1997) the provided anchor influenced the participants' estimates only when it was semantically related to the quantity to be estimated. Pohl (1998) found that the anchoring bias was absent when the anchor was perceived as implausible, and Hardt and Pohl (2003) found that the bias was smaller on trials where the anchor's judged plausibility was below the median plausibility judgment. Thus, at least under some circumstances, people appear to discard the provided value when it appears irrelevant or misleading.

However, realizing that the provided anchor is implausible and generating a better anchor require knowledge, effort, and time. Therefore, when people are asked to estimate a quantity they know almost nothing about, it may be resource-rational for them to anchor on whatever the experimenter suggested. This seems applicable to most anchoring experiments, because participants are usually so uncertain that they do not even know in which direction to adjust from the provided anchor (Simmons et al., 2010). If you cannot even tell whether the correct value is larger or smaller than the anchor, how could you generate a better one? The effect of the anchor is largest in people with little knowledge and high uncertainty about the quantity to be estimated (Jacowitz & Kahneman, 1995; Wilson et al., 1996). These people would benefit from a better anchor, but they cannot easily generate one, because they lack the relevant knowledge. Conversely, our simulation of the effect of knowledge suggested that people knowledgeable enough to generate good anchors, will perform well even if they start from a highly implausible anchor. Although this argument is speculative and has yet to be made precise it suggests that, at least in some situations, self-generating an anchor might not be worth the effort regardless of one's knowledge.

In conclusion, existing data are not necessarily inconsistent with the idea that anchors are chosen resource-rationally. Thus, whether anchors are chosen rationally is still an open question. Experimental and theoretical approaches to this question are an interesting avenue for future research that we will discuss below.

Resource-rationality: A better normative standard for human cognition?

When people estimate probabilities, the anchoring bias and other cognitive biases can cause their judgments to violate the laws of probability. This could be interpreted as a sign of human irrationality. However, adherence to the laws of logic and probability is just one of many notions of rationality. Existing definitions of rationality differ along four dimensions: The first distinction is whether rationality is defined in terms of beliefs (theoretical rationality) or actions (practical rationality, Harman, 2013; Sosis & Bishop, 2014). The second distinction is whether rationality is judged by the reasoning process or its outcome (Simon, 1976). Third, some notions of rationality take into account that the agent's computational capacity is bounded whereas others do not (Lewis et al., 2014; Russell, 1997). Fourth, rationality may be defined either by the agent's performance on a specific task or by its average performance in its natural environment (ecological rationality, Chater & Oaksford, 2014; Gigerenzer, 2008; Lewis et al., 2000).

In this taxonomy, Tversky and Kahneman's notion of rationality can be classified as theoretical, process-based, unbounded, and task-specific rationality. It is a notion of theoretical rationality, because it evaluates beliefs rather than actions. It is a form of process rationality, because it evaluates people by *how* they reason; specifically by whether or not their thoughts follow the rules of logic and probability theory. It is a notion of rationality for unbounded agents because it ignores the computational complexity of logical and probabilistic inference (Van Rooij, 2008). It is task-specific because it evaluates human rationality by people's performance on laboratory tasks specifically designed to elicit errors rather than representative everyday reasoning. We have argued that this is an unsuitable metric of human rationality and proposed a concrete alternative: resource-rationality. Resource-rationality differs from classical rationality along three of the four dimensions: First, it evaluates reasoning by its utility for subsequent decisions rather than by its formal correctness; this makes it an instance of practical rather than theoretical rationality. For instance, we evaluated anchoring-and-adjustment not by the correctness of the resulting estimates but by the rewards that people earned by using those estimates. Second, it agrees with Tversky and Kahneman's approach in that resource-rationality is an attribute of the process that generates conclusions and decisions. Third, it takes into account the cost of time and the boundedness of people's cognitive resources. Fourth, resource-rationality is defined with respect to the distribution of problem's in the agent's environment rather than a set of arbitrary laboratory tasks. Arguably, all three of these changes are necessary to obtain a normative–yet realistic–theory of human rationality. This new metric of rationality allowed us to re-evaluate the anchoring bias as a consequence of resource-rational computation rather than irrationality. Heuristics and rational models are often seen as opposites, but once the cost of computation is taken into account heuristics can be resource-rational. This illustrates the potential of resource-rational analysis to reconcile cognitive biases, such as the anchoring bias, with the fascinating capacities of human intelligence, and to connect rational theories, such as Bayesian models of cognition and rational analysis, to heuristics and other psychological process models (Griffiths et al., 2015).

Resource-rational analysis is closely related to other theoretical frameworks for analyzing cognition. The most closely related one is the *computational rationality* approach proposed by Lewis et al. (2014), which draws the same inspiration from Russell's work but focuses on finding optimal algorithms within a fixed cognitive architecture. Anderson's framework of *rational analysis* (1990, 1991), is also part of the inspiration of resource-rationality, although it provides only minimal treatment of the computational constraints under which organisms operate. Finally, the idea that human cognition is based on simple heuristics (Tversky and Kahneman, 1974; Gigerenzer & Selten, 2002) is compatible with resource-rationality – trading off errors with the cost of computation is exactly what good heuristics do. However, far from interpreting the cognitive biases resulting from such heuristics as evidence for human irrationality (Kahneman & Tversky, 1972; Nisbett & Borgida, 1975; Slovic et al., 1977) resource-rational analysis assumes that these biases are simply the consequence of rational use of limited computational resources.

Even though resource-rationality is a very recent approach, it has already shed some light on a wide range of cognitive abilities and provides a unifying framework for the study of intelligence in psychology, neuroscience, and artificial intelligence (Gershman et al., 2015). For example, we have recently applied the resource-rational framework to decision-making (Lieder et al., 2014), planning (Lieder et al., 2013), and strategy selection (Lieder et al., 2014; Lieder & Griffiths, 2015). In conclusion, resource-rationality appears to be a promising framework for normative and descriptive theories of human cognition.

Directions for future research

In a companion paper (Lieder, Griffiths, Huys, & Goodman, 2017) , we empirically confirm our model's prediction that adjustment increases with error cost but decreases with time cost. We show that this is true regardless of whether the anchor was provided or self-generated. This confirms our simulations' assumption that participants in numerical estimation experiments with provided anchors use the same cognitive strategy as participants in numerical estimation experiments with self-generated anchors.

The question to which extent anchors are chosen resource-rationally is one interesting avenue for future research. The hypothesis that anchors are chosen rationally predicts that if everything else is equal, then people will choose a relevant anchor over an irrelevant one. This could be probed by providing people with two anchors rather than just one. Alternatively, one could manipulate the ease of self-generating a good anchor and test whether this ease decreases the bias towards an implausible provided anchor. To analyze such experiments, the models developed could be used to infer which anchor people were using from the pattern of their responses.

An additional direction for future work is to extend the resource-rational anchoring-and-adjustment model. This could be done in several ways. First, the model could be extended by mechanisms for choosing and generating anchors. Second, the model could be extended by specifying *how* the mind approximates optimal resource allocation. A third extension of our model might incorporate directional information into the proposal distribution as in the Hamiltonian Monte Carlo algorithm (Neal, 2011) to better capture the effects of direction uncertainty discovered by Simmons et al. (2010). A fourth extension might capture the sequential incorporation of relevant knowledge by iterative conditioning and explore its connection to the selective accessibility theory of the anchoring bias (Strack and Mussweiler, 1997). A fifth frontier is to make resource-rational anchoring-and-adjustment more adaptive: How can the proposal distribution and a mechanism for choosing the number of adjustments be learned from experience? Can better performance be achieved by adapting the proposal distribution from one adjustment to the next? Finally, our resource-rational anchoring-and-adjustment model only uses a single sample, but it can be generalized to using multiple samples. Each of these extensions might improve the performance of the estimation strategy and it is an interesting question whether or not those improvements would bring its predictions closer to human behavior. Future studies might also evaluate additional alternatives to our model, such as an anchoring model with adaptive plausibility threshold or algorithms that directly approximate the most probable estimate rather than a sample from the posterior distribution.

Most previous models of heuristics are formulated for the domain in which the corresponding bias was discovered. For instance, previous models of anchoring-and-adjustment were specific to numerical estimation (Epley & Gilovich, 2006; Simmons et al., 2010). Yet, everyday reasoning is not restricted to numerical estimation and anchoring also occurs in very different domains such as social cognition (Epley et al., 2004). This highlights the challenge that models of cognition should be able to explain not only what people do in the laboratory but also their performance in the real-world. Heuristics should therefore be able to operate on the complex, high-dimensional semantic representations people use in everyday reasoning. Resource-rational anchoring-and-adjustment lives up to this challenge, because Markov-chain Monte Carlo methods are as applicable to semantic networks (Bourgin et al., 2014; Abbott et al., 2015) as they are to single numbers. In fact, resource-rational anchoring-and-adjustment is a very general mechanism that can operate over arbitrarily complex representations and might be deployed not only for numerical estimation but also in many other cognitive faculties such as memory retrieval, language understanding, social cognition, and creativity. For instance, resource-rational anchoring-and-adjustment may be able to explain the hindsight bias in memory recall (Pohl, 1998; Hardt & Pohl, 2003), primacy effects in sequential learning (Abbott & Griffiths, 2011), and the dynamics of memory retrieval (Abbott et al., 2015; Bourgin et al., 2014).

Conclusion

Resource-rational anchoring-and-adjustment provides a unifying, parsimonious, and principled explanation for a plethora of anchoring effects including some that were previously assumed to be incompatible with anchoring-and-adjustment. Interestingly, we discovered this cognitive strategy purely by applying resource-rational analysis to the problem of estimation under uncertainty. It is remarkable that the resulting model is so similar to the anchoring-and-adjustment heuristic. Our simulations support the conclusion that people rationally adapt the number of adjustments to the environment's incentives for speed and accuracy. Resource-rational anchoring and adjustment thereby reconciles the anchoring bias with people's adaptive intelligence and Bayesian models of reasoning under uncertainty. Concretely, the anchoring bias may reflect the optimal speed-accuracy tradeoff when errors are benign, which is true of most, if not all, laboratory tasks. Yet, when accuracy is important and speed is not crucial, then people perform more adjustments and the anchoring bias decreases. Hence, while people's estimates are biased in the statistical sense of the word ($\mathbb{E}\left[\hat{X}|K\right] \neq \mathbb{E}\left[X|K\right]$), our theory suggests that this is consistent with how they ought to reason. In this sense, the anchoring "bias" might not be a cognitive bias after all. Instead, the anchoring bias may be a window on resource-rational computation rather than a sign of human irrationality. Being biased can be resource-rational, and heuristics can be discovered by resource-rational analysis.

Acknowledgements This research was supported by grant number ONR MURI N00014-13-1-0341 from the Office of Naval Research (TLG and NDG), grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research (TLG), and a John S. McDonnell Scholar Award (NDG).

Appendix A

Notation

X:	numerical quantity to be estimated
$\hat{x}$:	people's estimates of quantity X
t:	number of adjustments
$\hat{x}_t$:	people's estimates of quantity X after t adjustments
K or y:	knowledge or information about X
$P(X\|K)$, $P(X\|y)$:	posterior belief about X
$P(R\|y)$:	distribution of people's responses to observation y
$\text{cost}(\hat{x}, x)$:	error cost of reporting estimate $\hat{x}$ when the true value is x
$t^\star$:	resource-rational number of adjustments
γ:	relative time cost per iteration
c_e, c_t:	cost of time, cost of error
Q:	approximate posterior belief
$\mathcal{H}$:	hypothesis space
μ_{prop}:	average size of proposed adjustments
$\mu^\star_{\text{prop}}$:	resource-rational step-size of proposed adjustments
a:	anchor

Appendix B

Generalization of optimal speed-accuracy tradeoff from problems to environments

Together, a person's knowledge K about a quantity X, the cost function $\text{cost}(\hat{x}, x)$, and the correct value x define an estimation problem. However, in most environments people are faced with many different estimation problems rather than just a single one, and the true values are unknown. We therefore define a task environment E by the relative frequency $P(X, K, \text{cost}|E)$ with which different estimation problems occur in it. Within each of the experiments that we are going to simulate, the utilities, and the participant's knowledge are constant. Thus, those task environments are fully characterized by $P(X, K|E)$ and $\text{cost}(\hat{x}, x)$.

The optimal speed-accuracy tradeoff weights the costs in different estimation problems according to their prevalence in the agent's environment. Formally, the agent should minimize the expected error cost in Eq. 2 with respect to the distribution of estimation problems $P(X, K|E)$ in its environment E:

$$t^\star = \arg\max_t \mathbb{E}_{P(X,K|E)}\left[\mathbb{E}_{Q(\hat{x}_t|K)}\left[u(x, \hat{x}_t) - \gamma \cdot t\right]\right]. \quad (6)$$

Thus, the number of adjustments is chosen to optimize the agent's average reward rate across the problem distribution of the task environment (cf. Lewis et al., 2014). If the task environment is an experiment with multiple questions, then the expected value is the average across those questions.

Appendix C

Estimating beliefs

For each simulated experiment we conducted one short online survey for each quantity X that its participants were asked to estimate. For each survey we recruited 30 participants on Amazon Mechanical Turk and asked the four questions that Speirs-Bridge et al. (2010) advocate for the elicitation of subjective confidence intervals: "Realistically, what do you think is the lowest value that the ... could be?", "Realistically, what do you think is the highest value that

Table 2 Estimated Beliefs: Insufficient adjustment from provided anchors

Study	Quantity	μ	σ	Correct
Tversky and Kahneman (1974)	African countries in UN (in %)	22.5	11.12	28
Jacowitz and Kahneman (1995)	length of Mississippi River (in miles)	1,525	770	2,320
Jacowitz and Kahneman (1995)	height of mount Everest (in feet)	27,500	3,902	29,029
Jacowitz and Kahneman (1995)	amount of meet eaten by average American (in pounds)	238	210	220
Jacowitz and Kahneman (1995)	distance from San Francisco to New York (in miles)	3000	718	2,900
Jacowitz and Kahneman (1995)	height of tallest redwood tree (in feet)	325	278	379.3
Jacowitz and Kahneman (1995)	number of United Nations members	111	46	193
Jacowitz and Kahneman (1995)	number of female professors at the University of California, Berkeley	83	251	805
Jacowitzx and Kahneman (1995)	population of Chicago (in millions)	5	3	2.715
Jacowitz and Kahneman (1995)	year telephone was invented	1885	35	1876
Jacowitz and Kahneman (1995)	average number of babies born per day in the United States	8,750	15,916	3,952,841
Jacowitz and Kahneman (1995)	maximum speed of house cat (in mph)	17	10	29.8
Jacowitz and Kahneman (1995)	amount of gas used per month by average American (in gallons)	55	84	35.2
Jacowitz and Kahneman (1995)	number of bars in Berkeley, CA	43	55	101
Jacowitz and Kahneman (1995)	number of state colleges and universities in California	57	112	248
Jacowitz and Kahneman (1995)	number of Lincoln's presidency	6	2	16

Table 3 Estimated beliefs: Insufficient Adjustment from self-generated anchors

Study by Epley, & Gilovich (2006)	Quantity	Mean	SD	Correct
Study 1a	Washington's election year	1786.5	7.69	1789
Study 1a	Boiling Point on Mount Everest in °F	158.8	36.82	160
Study 1a	Freezing Point of vodka in °F	3.7	17.052	−20
Study 1a	lowest recorded human body temperature in °F	86	14.83	55.4
Study 1a	highest recorded human body temperature in °F	108	3.39	115.7
Study 1b	Washington's election year	1786.5	7.69	1789
Study 1b	Boiling point in Denver in °F	201.3	9.93	203
Study 1b	Number of US states in 1880	33.5	8.52	38
Study 1b	year 2nd European explorer reached West Indies	1533.3	33.93	1501
Study 1b	Freezing point of vodka in °F	3.7	17.05	–20

Table 4 Estimated beliefs: Effect of cognitive load

Study by Epley, & Gilovich (2006)	Quantity	Mean	SD	Correct
Study 2b	Washington's election year	1786.5	7.69	1789
Study 2b	second explorer	1533.3	33.93	1501
Study 2c	Washington's election year	1786.5	7.69	1789
Study 2c	second explorer	1533.3	33.93	1501
Study 2c	Highest body temperature	108	3.39	115.7
Study 2c	boiling point on Mt. Everest	158.8	36.82	160
Study 2c	Lowest body temperature	86	14.83	55.4
Study 2c	freezing point of vodka	3.7	17.05	–20
Study 2c	number of U.S. states in 1880	33.5	8.52	38

Table 5 Estimated beliefs: effects of distance and knowledge

Study	Quantity	Mean	SD	Correct
Russo and Shoemaker (1989)	year of Atilla's defeat	953.5	398.42	451
Wilson et al. (1996); less knowledgeable group	Number of countries in the world	46.25	45.18	196
Wilson et al. (1996); knowledgeable group	Number of countries in the world	185	35.11	196

the ... could be?", "Realistically, what is your best guess (i.e. most likely estimate) of the ... ?", and "How confident are you that your interval from the lowest to the highest value could contain the true value o the ... ? Please enter a number between 0 and 100%.". These questions elicit a lower bound (l_s) and an upper bound (h_s) on the value of X, an estimate (m_s), and the subjective probability p_s that X lies between the lower and the upper bound ($P(X \in [l_s, h_s]|K)$ respectively, for each participant s. To estimate people's knowledge about each quantity from the reported confidence intervals, we modeled their belief $P(X|K)$ by a normal distribution $\mathcal{N}(\mu_s, \sigma_s)$. We used the empirical estimate m_s as μ_s, and set σ_s to $\frac{h_s - l_s}{\Phi^{-1}((1+p_s)/2) - \Phi^{-1}(1-(p_s+1)/2)}$, where Φ is the cumulative distribution function of the standard normal distribution. Finally, we took the medians of these estimates as the values of μ and σ used in our simulations. We applied this procedure separately for each quantity from each experiment that will be simulated below. The quantities and the estimated beliefs are summarized in Appendix C.

The hypothesis space $\mathcal{H}$ for each quantity was assumed to contain all evenly spaced values (interval $= \frac{\sigma}{20}$) in the range spanned by the 0.5th and the 99.5th percentile of the belief distribution $P(X|K)$ and the anchor(s) plus or minus one standard deviation. We simulated the adjustments people consider by samples from a Poisson distribution, that is $P(\delta = h_k - h_j) = \text{Poisson}(|k - j|; \mu_{\text{prop}})$, where h_k and h_j are the k^{th} and the j^{th} value in the hypothesis space $\mathcal{H}$, and μ_{prop} is the expected step-size of the proposal distribution $P(\delta)$. This captures the intuition that people consider only a finite number of discrete hypotheses and that the adjustments a person will consider have a characteristic size that depends on the resolution of her hypothesis space.

The following tables summarize our estimates of people's beliefs about the quantities used in the simulated anchoring experiments. Since the estimated probabilistic

Table 6 Estimated beliefs: Anchor type moderates effect of accuracy motivation; Abbreviations: EG– Epley & Gilovich (2005), TK– Tversky & Kahneman (1974)

Study	Quantity	Mean	SD	Correct
EG, Study 1	population of Chicago	5,000,000	2,995,797.04	2,719,000
EG, Study 1	height of tallest redwood tree	200	76.58	379.3
EG, Study 1	length of Mississippi river (in miles)	1875	594.88	2,320
EG, Study 1	height of Mt. Everest (in feet)	15400	4657.90	29,029
EG, Study 1	Washington's election year	1788	6.77	1789
EG, Study 1	year the 2nd explorer after Columbus reached the West Indies	1507.75	34.34	1501
EG, Study 1	boiling point on Everest (in °F)	150.25	36.82	160
EG, Study 1	freezing point of vodka (in °F)	−1.25	14.73	−20
EG, Study 2	Washington election year	1788	6.77	1789
EG, Study 2	2nd explorer	1507.75	34.34	1501
EG, Study 2	boiling point on Mt. Everest (in °F)	150.25	36.82	160
EG, Study 2	number of US states in 1880	33.5	8.52	38
EG, Study 2	freezing point of vodka (in °F)	−1.25	14.73	−20
EG, Study 2	population of Chicago	3000000	1257981.51	2,719,000
EG, Study 2	height of tallest redwood tree (in feet)	200	76.58	379.3
EG, Study 2	length of Mississippi river (in miles)	1875	594.88	2320
EG, Study 2	height of Mt. Everest	15400	4657.90	29,029
EG, Study 2	invention of telephone	1870	54.48	1876
EG, Study 2	babies born in US per day	7875	8118.58	3,952,841
TK	African countries in UN	22.5	11.12	28

Table 7 Estimated beliefs: Effects of direction uncertainty

Simmons et al. (2010)	**Quantity**	**Mean**	**SD**	**Correct**
Study 2	length of Mississippi river (in miles)	1625	752.3	2,320
Study 2	average annual rainfall in Philadelphia (in inches)	36.5	23.80	41
Study 2	Polk's election year	1857.5	45.42	1845
Study 2	Maximum speed of a house cat (miles per hour)	16	9.40	30
Study 2	Avg. annual temperature in Phoenix (in °F)	82.75	13.82	73
Study 2	Population of Chicago	2,700,000	1,560,608	2,719,000
Study 2	Height of Mount Everest (in feet)	23,750	7,519.70	29,032
Study 2	Avg. lifespan of a bullfrog (in years)	5.75	6.68	16
Study 2	Number of countries in the world	216.25	77.21	192
Study 2	Distance between San Francisco and Kansas city (in miles)	1,425	547.86	1,800
Study 3b	Year Seinfeld first aired	1991	2.23	1989
Study 3b	Average temperature in Boston in January	26.5	14.86	36
Study 3b	Year JFK began his term as U.S. president	1961.25	2.26	1961
Study 3b	Avg. temperature in Phoenix in Aug.	96	10.21	105
Study 3b	Year Back to the Future appeared in theaters	1985	1.54	1985
Study 3b	Avg. temperature in NY in Sept.	70	10.51	74

beliefs are normal distributions, we summarize each of them by a mean μ and a standard deviation σ.

References

Abbott, J.T., Austerweil, J.L., & Griffiths, T.L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569.

Abbott, J.T., & Griffiths, T.L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning *In Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, Texas: Cognitive Science Society.

Anderson, J.R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261–295.

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Psychology Press.

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–485.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). Coherent arbitrariness: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106.

Beach, L.R., & Mitchell, T.R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*(3), 439–449.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T.L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E., Denison, S., Griffiths, T.L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500.

Bourgin, D.D., Abbott, J.T., Griffiths, T.L., Smith, K.A., & Vul, E. (2014). Empirical evidence for markov chain monte carlo in memory search. *In Proceedings of the 36th annual meeting of the cognitive science society*, (pp. 224–229).

Braine, M.D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*(1), 1.

Brewer, N.T., & Chapman, G.B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making*, *15*, 65–77.

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, *7*(11), e1002211.

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

Chapman, G.B., & Johnson, E.J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In Gilovich, T., Griffin, D., & Kahneman, D. (Eds.) *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, U.K.: Cambridge University Press.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*(1), 93–131.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, *126*(2), 285–300.

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135.

Doucet, A., De Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200.

Epley, N. (2004). A tale of tuned decks?Anchoring as accessibility and anchoring as adjustment. In Koehler, D.J., & Harvey, N. (Eds.) *The Blackwell Handbook of Judgment and Decision Making* (pp. 240–256). Oxford, UK: Blackwell.

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin*, *30*(4), 447–460.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199–212.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. *Psychological Science*, *17*(4), 311–318.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Friedman, M., & Savage, L.J. (1948). The utility analysis of choices involving risk. *The Journal of Political Economy*, 279–304.

Friston, K. (2009). The free-energy principle: A rough guide to the brain?. *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Galinsky, A.D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, *81*(4), 657.

Gershman, S.J., Horvitz, E.J., & Tenenbaum, J.B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gershman, S.J., Vul, E., & Tenenbaum, J.B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., & Selten, R. (2002). In Gigerenzer, G., & Selten, R. (Eds.) *Bounded rationality: The adaptive toolbox*. Cambridge, MA: The MIT Press.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Good, I.J. (1983). *Good thinking: The foundations of probability and its applications*. USA: Univ Of Minnesota Press.

Griffiths, T.L., Lieder, F., & Goodman, N.D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Griffiths, T.L., & Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Griffiths, T.L., & Tenenbaum, J.B. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*(4), 725–743.

Habenschuss, S., Jonke, Z., & Maass, W. (2013). Stochastic computations in cortical microcircuit models. *PLoS Computational Biology*, *9*(11), e1003311.

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

Harman, G. (2013). Rationality. In LaFollette, H., Deigh, J., & Stroud, S. (Eds.) *International Encyclopedia of Ethics*. Hoboken: Blackwell Publishing Ltd.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hedström, P., & Stern, C. (2008). Rational choice and sociology. In Durlauf, S., & Blume, L. (Eds.) *The New Palgrave Dictionary of Economics*. 2nd edn. Basingstoke, U.K.: Palgrave Macmillan.

Horvitz, E., Suermondt, H., & Cooper, G. (1989). Bounded conditioning: Flexible inference for decisions under scarce resources *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence* (pp. 182–193). Mountain View: Association for Uncertainty in Artificial Intelligence.

Jacowitz, K.E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Lewis, R.L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311.

Lieder, F., Goodman, N.D., & Huys, Q.J.M. (2013). Controllability and resource-rational planning. In Pillow, J., Rust, N., Cohen, M., & Latham, P. (Eds.) *Cosyne Abstracts*.

Lieder, F., & Griffiths, T.L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In Noelle, D.C., et al. (Eds.) *Proceedings of the 37th Annual Conference of the Cognitive Science Society Austin*. TX: Cognitive Science Society.

Lieder, F., Griffiths, T.L., & Goodman, N.D. (2012). Burn-in, bias, and the rationality of anchoring. In Bartlett, P., Pereira, F.C.N., Bottou, L., Burges, C.J.C., & Weinberger, K.Q. (Eds.) *Advances in Neural Information Processing Systems 26*.

Lieder, F., Griffiths, T.L., Huys, Q.J.M., & Goodman, N.D. (2017). Empirical evidence for resource-rational anchoring-and-adjustment.

Lieder, F., Hsu, M., & Griffiths, T.L. (2014). The high availability of extreme events serves resource-rational decision-making. *In Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Lieder, F., Plunkett, D., Hamrick, J.B., Russell, S.J., Hay, N.J., & Griffiths, T.L. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. *Advances in Neural Information Processing Systems* 27.

Lohmann, S. (2008). Rational choice and political science. In Durlauf, S., & Blume, L. (Eds.) *The New Palgrave Dictionary of Economics*. 2nd edn. Basingstoke, U.K.: Palgrave Macmillan.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. W. H. Freeman. Paperback.

McKenzie, C.R. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and bayesian inference. *Cognitive Psychology*, *26*(3), 209–239.

Mengersen, K.L., & Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, *24*(1), 101–121.

Mill, J.S. (1882). *A system of logic ratiocinative and inductive*, 8th edn. New York: Harper and Brothers.

Moreno-Bote, R., Knill, D.C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(30), 12491–12496.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*(2), 136–164.

Neal, R. (2011). Brooks, S., Gelman, A., Jones, G., & Meng, X.L. (Eds.) *MCMC using Hamiltonian dynamics* (Vol. 2, pp. 113–162). FL, USA: CRC Press.

Neely, J.H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: General*, *106*(3), 226.

Newell, A., Shaw, J.C., & Simon, H.A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166.

Nisbett, R.E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, *32*(5), 932–943.

Nisbett, R.E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs: Prentice-Hall.

Northcraft, G.B., & Neale, M.A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, *39*(1), 84–97.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning (Oxford cognitive science series)*, 1st edn. Oxford: Oxford University Press.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1993). *The adaptive decision maker*: Cambridge University Press.

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

Russell, S.J. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*(1-2), 57–77.

Russell, S.J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Articial Intelligence Research*, *2*, 575–609.

Russell, S.J., & Wefald, E. (1991). *Do the right thing: Studies in limited rationality*. Cambridge, MA: The MIT Press.

Russo, J.E., & Schoemaker, P.J.H. (1989). *Decision traps: Ten barriers to brilliant decision-making and how to overcome them*: Simon and Schuster.

Sanborn, A.N., Griffiths, T.L., & Navarro, D.J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Schwarz, N. (2014). *Cognition and communication: Judgmental biases, research methods and the logic of conversation*. New York: Psychology Press.

Shafir, E., & LeBoeuf, R.A. (2002). Rationality. *Annual Review of Psychology*, *53*(1), 491–517.

Shugan, S.M. (1980). The cost of thinking. *Journal of consumer Research*, *7*(2), 99–111.

Simmons, J.P., LeBoeuf, R.A., & Nelson, L.D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*(6), 917–932.

Simon, H.A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Simon, H.A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*, 161–176.

Simon, H.A. (1976). From substantive to procedural rationality. In Kastelein, T.J., Kuipers, S.K., Nijenhuis, W.A., & Wagenaar, G.R. (Eds.) *25 Years of Economic Theory* (pp. 65–86). US: Springer.

Simonson, I., & Drolet, A. (2004). Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of Consumer Research*, *31*(3), 681–690.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Cognitive processes and societal risk taking. In Jungermann, H., & De Zeeuw, G. (Eds.) *Decision Making and Change in Human Affairs*, (Vol. 16 pp. 7–36). Dordrecht, Netherlands: D. Reidel Publishing Company.

Sosis, C., & Bishop, M. (2014). Rationality. *Wiley interdisciplinary reviews: Cognitive Science*, *5*, 27–37.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., & Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, *30*(3), 512–523.

Stewart, N., Chater, N., & Brown, G.D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

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

Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Computational Biology*, *9*(1), e1002803.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science*, *25*(3), 219–225.

Turner, B.M., & Schley, D.R. (2016). The anchor integration model: A descriptive model of anchoring effects. *Cognitive Psychology*, *90*, 1–47.

Turner, B.M., & Sederberg, P.B. (2012). Approximate bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(5), 375–385.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*(6), 939–984.

Von Neumann, J., & Morgenstern, O. (1944). *The theory of games and economic behavior*. Princeton: Princeton university press.

Vul, E., Goodman, N.D., Griffiths, T.L., & Tenenbaum, J.B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wilson, T.D., Houston, C.E., Etling, K.M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *125*(4), 387.

Wright, W.F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes*, *44*(1), 68–82.

Zhang, Y.C., & Schwarz, N. (2013). The power of precise numbers: A conversational logic analysis. *Journal of Experimental Social Psychology*, *49*(5), 944–946.