Strategy selection as rational metareasoning

Falk Lieder

Helen Wills Neuroscience Institute
University of California, Berkeley

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley

Author Note

This research was supported by grant number ONR MURI N00014-13-1-0341 from the Office of Naval Research to Thomas L. Griffiths. The source code of our model, simulations, and experiments are available on the Open Science Framework (https://osf.io/7yqbh/). The data are available upon request. A preliminary version of the model and its evaluation in the domain of sorting (Experiment 1) was presented at the 28th Annual Conference on Neural Information Processing Systems and appeared in the conference proceedings (Lieder et al., 2014). Preliminary versions of the simulation of Payne et al. (1998), the simulation of Rieskamp and Otto (2006) were presented at the 37th Annual Meeting of the Cognitive Science Society and appeared in the conference proceedings (Lieder & Griffiths, 2015). This work has been updated and extended for inclusion in the current manuscript.

Abstract

Many contemporary accounts of human reasoning assume that the mind is equipped with multiple heuristics that could be deployed to perform a given task. This raises the question of how the mind determines when to use which heuristic. To answer this question, we developed a rational model of strategy selection, based on the theory of rational metareasoning developed in the artificial intelligence literature. According to our model people learn to efficiently choose the strategy with the best cost-benefit tradeoff by learning a predictive model of each strategy's performance. We found that our model can provide a unifying explanation for classic findings from domains ranging from decision-making to arithmetic by capturing the variability of people's strategy choices, their dependence on task and context, and their development over time. Systematic model comparisons supported our theory, and four new experiments confirmed its distinctive predictions. Our findings suggest that people gradually learn to make increasingly more rational use of fallible heuristics. This perspective reconciles the two poles of the debate about human rationality by integrating heuristics and biases with learning and rationality.

Keywords: bounded rationality; strategy selection; heuristics; meta-decision-making; metacognitive reinforcement learning

To succeed in life we have to solve a wide range of problems that place very different demands on us: sometimes we have to think fast and sometimes we have to think slow (cf. Kahneman, 2011). For instance, avoiding a car accident requires a split-second decision, whereas founding a successful start-up requires investing a lot of time into anticipating the future and weighting potential outcomes appropriately. No single decision mechanism works well across all situations. To meet the wide range of demands posed by different decision problems, it has been proposed that the human brain is equipped with multiple decision systems (Dolan & Dayan, 2013) and decision strategies (Payne, Bettman, & Johnson, 1988). Dual-process theories are a prominent example of this perspective (Evans & Stanovich, 2013; Evans, 2003; Kahneman, 2011). The coexistence of multiple alternative strategies is not specific to decision making. People also appear to possess multiple strategies for inference (Gigerenzer & Selten, 2002), memory (Bjorklund & Douglas, 1997), self-control (Braver, 2012), problem solving (Fum & Del Missier, 2001), and mental arithmetic (Siegler, 1999) to name just a few.

The availability of multiple strategies that are applicable to the same problems raises the question how people decide when to use which strategy. The fact that so many different strategies have been observed under different circumstances shows that people's strategy choices are highly variable and contingent on the situation and the task (Beach & Mitchell, 1978; Fum & Del Missier, 2001; Payne, 1982; Payne et al., 1988). Overall, the contingency of people's strategy choices appears to be adaptive. Even though under certain circumstances people have been found to use heuristics that cause systematic errors (Ariely, 2009; Sutherland, 2013), their strategies are typically well-adapted to the problems to which they are applied (Anderson, 1990; Braver, 2012; Bröder, 2003; Fum & Del Missier, 2001; Payne, Bettman, & Johnson, 1993). For instance, Payne and colleagues found that when the probabilities of alternative outcomes fall off quickly, then decision makers employ frugal heuristics that prioritize the most probable outcomes at the expense of less probable ones. Similarly, decision makers select fast heuristics when they are under time pressure but more accurate ones when they are not (Payne et al., 1988). These and other studies (e.g. Siegler, 1999) have also documented that people's propensity to use one strategy rather than another changes over time.

The adaptiveness of people's strategy choices appears to increase with experience. For instance, as children gain more experience with mental arithmetic they gradually learn to choose effective and efficient strategies more frequently (Siegler, 1999). In adults, adaptive changes in strategy selection have been observed on much shorter time scales. For instance, adults have been found to adapt their decision strategy to the structure of their decision environment within minutes as they repeatedly choose between different investment based on multiple attributes (Rieskamp & Otto, 2006): In a decision environment where the better investment option is determined by a single attribute people learn to use a fast-and-frugal heuristic that ignores all other attributes. But when the decision environment does not have that structure, then people learn to integrate multiple attributes.

How can we explain the variability, task- and context-dependence, and change in people's strategy choices? Despite the previous work reviewed in the following section and some recent progress on how the brain decides how to decide (Boureau, Sokol-Hessner, & Daw, 2015) the strategy selection problem remains unsolved (Marewski & Link, 2014). Finally, while it is typically assumed that people's use of heuristics is irrational (Ariely, 2009; Marcus, 2009; Sutherland, 2013), there is increasing evidence for adaptive strategy selection (Boureau et al., 2015; Braver, 2012; Daw, Niv, & Dayan, 2005; Fum & Del Missier, 2001; Gunzelmann & Anderson, 2003; Keramati, Dezfouli, & Piray, 2011; Payne et al., 1988). This raises the additional question whether and to what extent people's strategy choices are rational.

In this article we formalize the strategy selection problem, derive a rational strategy selection mechanism, and show that it can explain a wide range of empirical phenomena including the variability, contingency, and change of strategy selection across multiple domains – ranging from decision-making to arithmetic – and time scales. Our theory adds an important missing piece to the puzzle of bounded rationality by specifying when people should use which heuristic, and our findings reconcile the two poles of the debate about human rationality by suggesting that people gradually learn to make increasingly more rational use of their fallible heuristics.

In the next section, we situate our work in the debate about human rationality and previous research on strategy selection. We then develop an alternative, rational account of strategy selection based on the idea of *rational metareasoning* from artificial intelligence research (Russell & Wefald, 1991). In the following sections, we evaluate our theory against traditional theories of strategy selection and show that it provides a unifying explanation for a wide range of phenomena: We show that rational metareasoning can account for people's ability to adaptively choose the sorting strategy that works best for each individual problem based on limited experience, while traditional theories of strategy selection cannot. In the subsequent sections, we show that this conclusion holds not only for behavioral strategies but is equally true of cognitive strategies for decision-making, and mental arithmetic that operate on internal representations. We conclude with the implications of our findings for the debate about human rationality and directions for future research.

Background

The debate about human rationality

Historically, rationality has been defined as reasoning according to the laws of logic and probability theory and making decisions that conform to the axioms of expected utility theory (Von Neumann & Morgenstern, 1944). Consequently, the debate whether people are rational has traditionally been focused on whether or not people's judgments and decisions follow the rules of these normative theories (Stanovich, 2009). Numerous studies suggested that human judgments systematically violate the laws of logic (e.g., Wason, 1968; Tversky & Kahneman, 1983) and probability theory (Tversky & Kahneman, 1974), and that our decisions fall short of the prescriptions of expected utility theory (Kahneman & Tversky, 1979). These cognitive biases have been shown to result from people's reliance on simple heuristics that sacrifice guarantees of optimality for speed and efficiency (Tversky & Kahneman, 1974).

Under the classical definition of rationality, it is irrational to rely on heuristics because they give rise to cognitive biases. Yet, this classical definition does not take into account that our decisions and judgments have to be made with limited cognitive resources in finite time. Hence, while the demonstration of cognitive biases suggest that we are not unboundedly rational, they do not rule out the possibility that people make rational use of their finite time and limited cognitive resources. This hypothesis can be traced back to Simon's notion of bounded rationality (Simon, 1955, 1972, 1956), but it has only recently been formalized mathematically (Griffiths, Lieder, & Goodman, 2015; Lewis, Howes, & Singh, 2014). In the following, we will refer to cognitive strategies that make optimal use of finite cognitive resources as *resource-rational*.

Recent studies have found that major cognitive biases in judgment and decision-making that have been interpreted as evidence against human rationality are consistent with the rational use of finite cognitive resources. Concretely, the anchoring bias that pervades human judgment appears to be the manifestation of a resource-rational strategy for drawing inferences under uncertainty (Lieder, Griffiths, & Goodman, 2012) and numerous cognitive biases in people's decisions under uncertainty are accurately predicted by a resource-rational decision strategy (Lieder, Hsu, & Griffiths, 2014; Lieder, Griffiths, & Hsu, in press). This line of work demonstrates that fallible heuristics can be resource-rational for certain problems under some circumstances. Similarly, Gigerenzer and colleagues have found that simple, fast-and-frugal heuristics perform very well when their assumptions match the structure of the environment (Gigerenzer, Todd, & The ABC Group, 1999; Gigerenzer & Brighton, 2009; Gigerenzer & Selten, 2002; Gigerenzer, 2008a, 2008b; Todd & Gigerenzer, 2012).

Scholars who view heuristics as irrational kluges that give rise to fallacies and biases (Ariely, 2009; Marcus, 2009; Sutherland, 2013) emphasize situations in which the chosen heuristics are maladaptive, whereas researchers who interpret heuristics as rational strategies point to situations where people use them adaptively (Todd & Gigerenzer, 2012; Griffiths, Lieder, & Goodman, 2015). Arguably, most heuristics are neither rational nor irrational per se. Instead, their rationality depends on how well they fit the problem to which they are being applied. Hence, the degree to which people are rational depends on when they use which heuristic. The critical question thus becomes "Are heuristics chosen rationally?" In this article, we address this question by developing and testing a rational model of strategy selection.

Previous theories of strategy selection

Strategy selection was initially viewed as a metacognitive decision based on explicit metacognitive knowledge about which cognitive strategies are best suited for which purposes (Flavell, 1979). Consistent with this perspective, Beach and Mitchell (1978) proposed that people choose decision strategies by performing an explicit cost-benefit analysis. Although Beach and Mitchell (1978) did not formalize this process enough to make quantitative predictions, their qualitative predictions were later confirmed in the domain of decision-making (Payne et al., 1988). Payne and colleagues demonstrated that which decision process performs best is contingent on time pressure and the structure of the decision problem.

The participants in the experiments conducted by Payne et al. (1988) responded adaptively to task contingencies *as if* their strategy choices were based on a cost-benefit analysis. Yet, under most circumstances, performing a complete cost-benefit analysis would take substantially longer than executing the most accurate strategy. In order to be beneficial, people's strategy selection mechanism has to be efficient. Furthermore, it has to avoid the infinite regress that could potentially result from reasoning about reasoning. These considerations have led researchers to abandon the idea that strategies are selected by a metacognitive cost-benefit analysis in favor of simpler models that select strategies by learning directly from experience (Erev & Barron, 2005; Rieskamp & Otto, 2006; Shrager & Siegler, 1998; Siegler & Shrager, 1984; Siegler, 1988). Consistent with this emphasis on learning, multiple experiments have found that people's strategy choices become more adaptive with experience (Bröder, 2003; Payne et al., 1988; Rieskamp & Otto, 2006).

Previous learning-based accounts of strategy selection were based on simple associative learning (Shrager & Siegler, 1998) and learning from feedback (Erev & Barron, 2005; Rieskamp & Otto, 2006). These mechanisms can be interpreted as a form of model-free metacognitive reinforcement learning in the sense they update the decision-maker's propensity to choose a strategy directly without building a model of what will happen when the strategy is selected[1]. According to the SSL (Rieskamp & Otto, 2006) and RELACS (Erev & Barron, 2005) models (defined in detail in Appendix A), people solve the strategy selection problem by learning which strategy works best on average in a given environment. This learning mechanism does not exploit the fact that every problem has distinct characteristics that determine the strategies' effectiveness.

According to the SCADS model (Shrager & Siegler, 1998), people learn to associate strategies with problem types. Every time a strategy is applied to a problem the association between the problem's type and the strategy is strengthened, and this strengthening is strongest when the strategy was successful. Using the same mechanism, the SCADS model also learns a global association between each strategy and problems in general. When presented with a problem the SCADS model chooses the strategy for which the product of the problem type specific association strengths and the global association strength is highest. This learning mechanism presupposes that each problem has been identified as an instance of one or more problem types. If each problem belongs to exactly one category, then the SCADS model learns to use the same strategy for all problems of a given type, but each problem can belong to multiple categories.

In his rational analysis of problem solving Anderson (1990) developed a more sophisticated strategy selection mechanism according to which people probabilistically select strategies (productions) that yield a high value of $\hat{P} \cdot G - \hat{C}$ where G is the value of achieving the goal and $\hat{P}$ and $\hat{C}$ are Bayesian estimates of the success probability and the cost of achieving the goal. This mechanism has been implemented in ACT-R to simulate strategy selection learning in problem solving (Gunzelmann & Anderson, 2003). However, like the model-free reinforcement learning mechanisms of SSL and RELACS (Erev & Barron, 2005; Rieskamp & Otto, 2006) the learning mechanism of ACT-R does not exploit the fat that some problems are more similar than others.

The cognitive niche theory (Marewski & Schooler, 2011) complements theories points out that people need only choose between those strategies that are applicable in a given situation. It emphasizes that the affordances of most situations severely limit the number of applicable strategies, for instance because the information required by many strategies is unavailable or cannot be recalled.

Recent work in computational neuroscience has modeled how the brain arbitrates between the model-free (habitual) and the model-based (goal-directed) decision system as

[1] From a different perspective, all theories of strategy selection learning can be seen as model-based because each strategy corresponds to a certain model of the environment (Gluth, Rieskamp, & Büchel, 2014).

meta-decision-making using ideas from reinforcement learning (Boureau et al., 2015; Daw et al., 2005; Keramati et al., 2011). This approach is promising and the reinforcement-learning framework is very powerful. However, it has yet to be extended to the complexities of the more general problem of strategy selection. In the following section, we pursue this idea to provide a new rational analysis of strategy selection that overcomes the limitations of previous theories.

Strategy selection learning as metacognitive reinforcement learning

In this section we provide a computational-level theory (Marr, 1982) of the strategy selection problem and propose a learning and a selection mechanism through which people might solve this problem. The key idea is that people learn to predict the accuracy and execution time of each strategy from features of individual problems and choose the strategy with the best predicted speed-accuracy tradeoff.

The strategy selection problem

Each environment E can be characterized by the relative frequency P_E with which different kinds of problems occur in it. In most environments, these problems are so diverse that none of people's strategies can achieve the optimal speed-accuracy tradeoff on all of them. Optimal performance in such environments requires selecting different strategies for different types of problems. One way to achieve this would be to learn the optimal strategy for each problem separately through trial and error. This approach is unlikely to succeed in complex environments where no problem is exactly the same as any of the previous ones. Hence, in many real-world environments, learning about each problem separately would leave the agent completely unprepared for problems it has never seen before. This can be avoided by exploiting the fact that each problem has perceivable features $f_1, \cdots, f_K$ that can be used to predict the performance of candidate strategies from their performance on previous problems. For instance, the features of a risky choice may include the number of options, the spread of the outcome probabilities, and the range of possible payoffs.

How good it is to apply strategy s to $\text{problem}^{(\text{i})}$ depends not only on the expected reward but also on the expected cost of executing the strategy. Building on the theory of rational metareasoning developed in artificial intelligence research (Russell & Wefald, 1991), this can be quantified by the value of computation (VOC):

$$\texttt{VOC}\left(s, \text{problem}^{(\text{i})}\right) = \mathbb{E}\left[U\left(s(\text{problem}^{(\text{i})}); \text{problem}^{(\text{i})}\right) - \text{cost}\left(s, \text{problem}^{(\text{i})}\right)\right],$$

where $s(\text{problem}^{(\text{i})})$ is the action the potentially stochastic strategy s selects on $\text{problem}^{(\text{i})}$, $U(A)$ denotes the utility of taking action A, and $\text{cost}\left(s, \text{problem}^{(\text{i})}\right)$ is the computational cost of executing strategy s on that problem. In the following we will assume that the computational cost is driven primarily by the (cognitive) opportunity cost of the strategy's execution time $T\left(s, \text{problem}^{(\text{i})}\right)$, that is

$$\text{cost}\left(s, \text{problem}^{(\text{i})}\right) = \gamma \cdot T\left(s, \text{problem}^{(\text{i})}\right).$$

The problem of optimal strategy selection can be defined as finding a mapping $m: \mathcal{F} \mapsto \mathcal{S}$ from feature vectors ($\boldsymbol{f}^{(i)} = (f_1(\text{problem}^{(\text{i})}), \cdots, f_K(\text{problem}^{(\text{i})})) \in \mathcal{F}$) to strategies ($s \in \mathcal{S}$) that maximizes the expected VOC of the selected strategy across all problems the environment might present. Hence, we can define the strategy selection problem as

$$\arg\max_{m} \sum_{\text{problem} \in \mathcal{P}} P_E(\text{problem}) \cdot \text{VOC}(m(\mathbf{f}(\text{problem})),\text{problem}),$$

where $\mathcal{P}$ is the set of problems that can occur.

Critically, the VOC of each strategy depends on the problem, but the strategy has to be chosen entirely based on the perceivable features $\boldsymbol{f}$ and the strategy selection mapping m has to be learned from experience. In machine learning, these kinds of problems are known as contextual multi-armed bandits (May, Korda, Lee, & Leslie, 2012). Two critical features of this class of problems are that they impose an exploration-exploitation tradeoff and require generalization. In the next section, we will leverage these insights to derive a rational strategy selection learning mechanism.

The experience gained from applying a strategy s to a problem with perceivable features $\boldsymbol{f}$ and observing an outcome with utility u after executing the strategy for t units of time can be summarized by the tuple $(\boldsymbol{f}, s, u, t)$. Hence, people's experience after the first n problems can be summarized by the history

$$h_n = \left(\left(\boldsymbol{f}^{(1)}, s^{(1)}, u^{(1)}, t^{(1)}\right), \cdots, \left(\boldsymbol{f}^{(n)}, s^{(n)}, u^{(n)}, t^{(n)}\right)\right),$$

where $\boldsymbol{f}^{(i)}, s^{(i)}, u^{(i)}$, and $t^{(i)}$ are the feature vector of the i^{th} problem, the strategy that was applied to it, and the resulting utility and execution time respectively. Strategy selection learning induces a sequence $m^{(1)}, m^{(2)}, \cdots, m^{(N)}$ of strategy selection mappings that depends on the agent's experience (h_n) and its strategy selection learning mechanism $l: \mathcal{H} \mapsto \mathcal{M}$ where $\mathcal{H}$ is the set of possible histories and $\mathcal{M}$ is the set of possible strategy selection mechanisms. With this notation, we can express the agent's performance on the n^{th} problem by

$$\text{VOC}\left(m^{(n)}\left(\boldsymbol{f}^{(n)}\right), \text{problem}^{(\text{n})}\right),$$

where $m^{(n)}\left(\boldsymbol{f}^{(n)}\right)$ is the strategy the agent selects for the n^{th} problem, and the strategy selection mapping $m^{(n)}$ is $l(h^{(n-1)})$. Since the problem is sampled at random, the expected performance at time step n is

$$V_n(l) = \mathbb{E}_{P_E}\left[\text{VOC}\left(m^{(n)}\left(\boldsymbol{f}^{(n)}\right), \text{problem}^{(\text{n})}\right) \mid m^{(n)} = l\left(h^{(n-1)}\right)\right].$$

If the agent solves N problems before it runs out of time, its total performance is

$$V_{\text{total}}(l) = \mathbb{E}\left[\sum_{n=1}^{N} V_n(l)\right].$$

Using this notation, we can define the optimal strategy selection learning mechanism $l^\star$ as the one that maximizes the agent's total expected value of computation across all possible sequences of problems, that is

$$l^\star = \arg\max_{l} V_{\text{total}}(l).$$

This concludes our computational-level analysis of strategy selection and strategy selection learning. We will now use this analysis as a starting point for deriving a rational strategy selection learning mechanism.

A rational process model of strategy selection

Our computational-level analysis identified that a general strategy selection learning mechanism should be able to transfer knowledge gained from solving one problem to new problems that are similar. In the reinforcement learning literature generalization is typically achieved by parametric function approximation (Sutton &

Barto, 1998). The simplest version of this approach is to learn the coefficients of a linear function predicting the value of a state from its features. Such linear approximations require minimal effort to evaluate and can be learned very efficiently. We therefore propose that people learn an internal predictive model that approximates the value of applying a strategy s to a problem by a weighted average of the problem's features $f_1(\text{problem}), \cdots, f_n(\text{problem})$:

$$\text{VOC}(s, \text{problem}) \approx \sum_{k=1}^{n} w_{k,s} \cdot f_k(\text{problem}). \qquad (1)$$

This approximation is easy to evaluate, but it is not clear how it can be learned given that the VOC cannot be observed directly. However, when the strategy s generates a decision, then the VOC can be decomposed into the uti1.lity of the decision's outcome and the cost of executing the strategy. Assuming that the cost of executing the strategy is proportional to its execution time, the VOC can be approximated by

$$\text{VOC}(s, \text{problem}) \approx \mathbb{E}[U|\text{ problem}, \mathrm{s}] - \gamma \cdot \mathbb{E}[T \mid \text{problem}, s], \qquad (2)$$

where U is the utility of the outcome obtained by following strategy s, γ is the agent's opportunity cost per unit time and $\mathbb{E}[T \mid \text{problem}, s]$ is the expected execution time of the strategy s when applied to the problem.

Approximating the VOC thus becomes a matter of estimating three quantities: the expected utility of relying on the strategy, the opportunity cost per unit time, and the expected time required to execute the strategy. The agent can learn its opportunity cost γ by estimating its reward rate (Boureau et al., 2015; Niv, Daw, Joel, & Dayan, 2007), and the utility of applying the strategy and its execution time T can be observed. Therefore, when the reward is continuous, then it is possible to learn an efficient approximation to the VOC by learning linear predictive models of the utility of its decisions and its execution time and combining them according to

$$\text{VOC}(s, \text{problem}) \approx \sum_{k=1}^{n} w_{k,s}^{(R)} \cdot f_k(\text{problem}) - \hat{\gamma} \cdot \sum_{k=1}^{n} w_{k,s}^{(T)} \cdot f_k(\text{problem}).$$

This equation is a special case of the general approach specified in Equation 1. When the outcome is binary, then the predictive model of the reward takes the form

$$P(O = 1|s, \text{problem}) = \frac{1}{1 + \exp\left(-\sum_{k=1}^{n} w_{k,s}^{(R)} \cdot f_k(\text{problem})\right)}.$$

We model the agent's estimate of its opportunity cost γ by the posterior mean $\mathbb{E}[\bar{r}|t_{\text{total}}, r_{\text{total}}]$ of its reward rate $\bar{r}$ given the sum of rewards r_{total} that the agent has experienced and the time since the beginning of the experiment (t_{total}). To do so, we assume that both the prior and the likelihood function are Gaussian, that is

$$P\left({}^{r_{\text{total}}}/_{t_{\text{total}}}\middle|\,\bar{r}\right) = \mathcal{N}\left(\mu = \bar{r}, \tau = t_{\text{total}} \cdot {}^{1}/_{60s}\right),$$
$$P(\bar{r}) = \mathcal{N}(1,1).$$

In this model, the weight of the agent's experience increases linearly with its time spent in the environment, and the prior corresponds to 60 sec worth of experience.

Our theory covers learning and strategy selection. To simulate learning, the agent's belief about the reward rate and the feature weights in the predictive model of a strategy's accuracy and execution time are updated by Bayesian learning every time it has been executed: The belief about the reward rate $\bar{r}$ is updated to $P(\bar{r}|\mathrm{r}_{\text{total}}, t_{\text{total}})$ as described in Section 2 of Appendix A. The weights of the execution time model are updated by Bayesian linear regression (see Section 3 of Appendix A). The weights of the reward model are updated by Bayesian logistic regression (see Section 4 of Appendix A) if the reward is binary (i.e., correct vs. incorrect), or by Bayesian linear regression (see Section 3 of Appendix A) when the reward is continuous (e.g., monetary). Lastly, our model learns which features are relevant for predicting the most recent strategy's execution time and reward by performing Bayesian model selection as described in Section 5 of Appendix A.

The second component of our model is strategy selection. Given the learned predictive models of execution time and reward, the agent could predict the expected VOC of each available strategy and select the strategy with the highest expected VOC. While this approach works well when the agent has already learned a good approximation to the VOC of each strategy, it ignores the value of learning about strategies whose performance is still uncertain. Hence, always using the strategy that appears best could prevent the agent from discovering that other strategies work even better. Yet, on average, strategies that appear sub-optimal will choose worse actions than the strategy that appears best. This problem recapitulates the well-known exploration-exploitation dilemma in reinforcement learning. To solve this problem our model selects strategies by Thompson sampling (May, Korda, Lee, & Leslie, 2012; Thompson, 1933): For each strategy s, our model samples estimates $\widetilde{w} = \left(\widetilde{w}_{k,s}^{(T)}, \widetilde{w}_{k,s}^{(R)}\right)$ of the weights $w = \left(w_{k,s}^{(T)}, w_{k,s}^{(R)}\right)$ of the predictive models of execution time and reward from their respective posterior distributions, that is

$$\widetilde{w}_{k,s}^{(T)} \sim P\left(w_{k,s}^{(T)} \middle| h_{t-1,s}\right),$$
$$\widetilde{w}_{k,s}^{(R)} \sim P\left(w_{k,s}^{(R)} \middle| h_{t-1,s}\right),$$

where $h_{t-1,s}$ is the agent's past experience with strategy s at the beginning of trial t. From these weights $\widetilde{w}$, our model predicts the VOC values of all strategies s by

$$\hat{V}_t(s, \text{problem}) = \sum_{k=1}^{n} \widetilde{w}_{k,s}^{(R)} \cdot f_k(\text{problem}) - \mathbb{E}[\hat{\gamma}|h_t] \cdot \sum_{k=1}^{n} \widetilde{w}_{k,s}^{(T)} \cdot f_k(\text{problem}),$$

where $\mathbb{E}[\hat{\gamma}|h_t]$ is the posterior expectation of the agent's reward rate given its past experience. Finally, our model selects the strategy $s_t^\star$ with the highest predicted VOC,

$$s_t^\star = \text{argmax}_s\, \hat{V}_t(s, \text{problem}).$$

This concludes the description of our model.

Our proposal is similar to model-based reinforcement learning (Dolan & Dayan, 2013; Gläscher, Daw, Dayan, & O'Doherty, 2010) in that it learns a predictive model. However, both the predictors and the predicted variables are different. While model-based reinforcement learning aims to predict the next state and reward from the agent's action (e.g., "Go left!"), our model learns to predict the costs and benefits of the agent's deliberation from the agent's cognitive strategy (e.g., planning four steps ahead vs. planning only one step ahead). While model-based reinforcement learning is about the control of behavior, our model is about the control of mental activities that may have no direct effect on behavior. In brief, the main difference is that we have modeled metacognitive learning instead of stimulus-response learning. Despite this difference in semantics, the proposed learning mechanism is structurally similar to the semi-gradient SARSA algorithm from the reinforcement learning literature (Sutton & Barto, 1998).

As illustrated in Figure 1, our model's prediction mechanism could be approximated by a simple feed-forward neural network: The first layer represents the input to the strategy selection network. The subsequent hidden layers extract features that are predictive of the strategy's execution time and accuracy. The second last layer computes a linear combination of those features to predict the execution time and external reward of applying the strategy, and the final layer combines these predictions into an estimate of the VOC of applying the strategy in the current state. The weights of this network could be learned by a basic error-driven learning mechanism, and the features might emerge from applying the same error-driven learning mechanism to connections between the hidden layers (cf. Mnih et al., 2015). With one such network per strategy a simple winner-take-all network (Maass, 2000) could read out the strategy with the highest VOC. This neural network formulation suggests that a single forward pass through a small number of layers may be sufficient to compute each strategy's VOC. The action potentials and synaptic transmission required to propagate neural activity from one layer to the next happens in milliseconds. The winner-take-all mechanism for reading out the strategy with the highest VOC can be performed in less than one tenth of a second (Oster, Douglas, & Liu, 2009). Hence, the brain might be able to execute the proposed strategy selection mechanism within fractions of a second.

Summary

We have derived a rational process model of strategy selection as an efficient approximation to the optimal solution prescribed by rational metareasoning. In contrast to previous accounts of strategy selection, our model postulates a more sophisticated, feature-based representation of the problem to be solved and a learning mechanism that achieves generalization. Instead of just learning about the reward that each strategy obtains *on average* our model learns to predict each strategy's execution time and expected reward on each individual problem. Hence, while previous models learned which strategy works best on average, our model learns to predict which strategy is best for each individual problem. Whereas previous theories of strategy selection (Erev & Barron, 2005; Rieskamp & Otto, 2006; Siegler & Shipley, 1995; Siegler & Shrager, 1984; Siegler, 1988) rejected the ideal of a cost-benefit analysis as intractable, we propose that people learn to approximate it efficiently. Note, however, that the consideration of the cost of thinking (Shugan, 1980) is not the distinguishing feature of our model because costs can be incorporated into the reward functions of previous theories of strategy selection. Rather, the main innovation of our theory is that strategies are chosen based on the features of the problem to be solved. In the remainder of the article we show that this allows our model to capture aspects of human cognition that were left unexplained by previous theories.

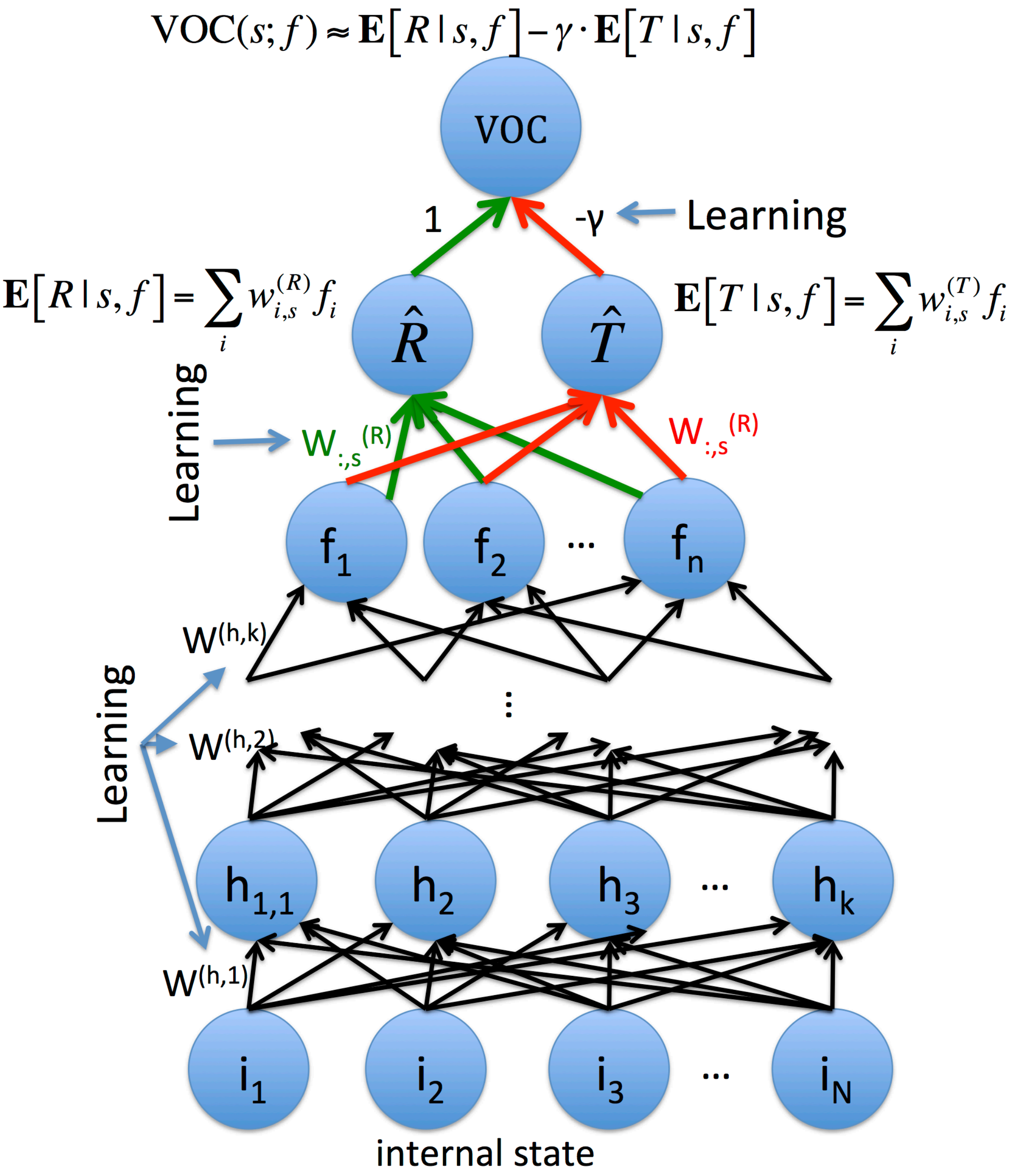


Figure 1: Our rational process model of strategy selection learning could be implemented in a simple feed-forward neural network.

Evaluating the model with sorting strategies

To test whether our rational model of strategy selection leaning can better account for how people's strategy choices change with experience than traditional context-free accounts, like RELACS, SSL and SCADS, we designed an experiment in which feature-based versus context-free strategy selection learning make qualitatively different predictions.[2] To differentiate between these two accounts we chose a domain in which the performance of alternative strategies is well understood and differs dramatically depending on easily detectable features of the problem. Furthermore, we were looking for a domain in which we could directly observe people's strategy choices. These considerations led us to study how people learn to choose between two alternative strategies for sorting a list of numbers: *cocktail sort* and *merge sort* (Knuth, 1998). We chose these two sorting strategies because they have opposite strengths and weaknesses. Cocktail sort is very fast for short and nearly-sorted lists, but in the worst case its runtime increases quadratically with the length of the list ($O(n^2)$). Thus. for long, unsorted, or reversely sorted lists cocktail sort is extremely inefficient. By contrast, the execution time of merge sort does not depend on the degree to which the list is correctly or reversely sorted and its execution time increases only log-linearly with the length of the list ($O(n \cdot \log(n))$). In the following we will assume that the task is to sort a list of numbers in ascending order.

To apply our theory to model how people learn to select between these two sorting strategies, we have to specify the features by which sorting problems are represented. For simplicity, we assume that the basic features are the length $|L|$ of the list $L = (e_1, e_2, \cdots, e_{|L|})$ and a measure of its presortedness:

$$f_1 = |L|,$$
$$f_2 = |\{m: e_m > e_{m+1}\}|,$$

where $|A|$ denotes the number of elements in the set or list A. The second feature estimates the number of pairs of elements that would have to be swapped in order to sort the list in ascending order from the number of times one element is larger than the next. Since it is well known that the runtimes of sorting algorithms are polynomials in the length of the list and its logarithm, we assume that the feature vector $\boldsymbol{f}$ includes all terms of the form

$$f_1^{k_1} \cdot \log(f_1)^{k_2} \cdot f_2^{k_3} \cdot \log(f_2)^{k_4},$$

where $k_1, k_2, k_3, k_4 \in \{0,1,2\}$ and $\sum_i k_i \leq 2$. As described above, our model will select a subset of these features and use them to predict the execution time and success probability of each sorting strategy.

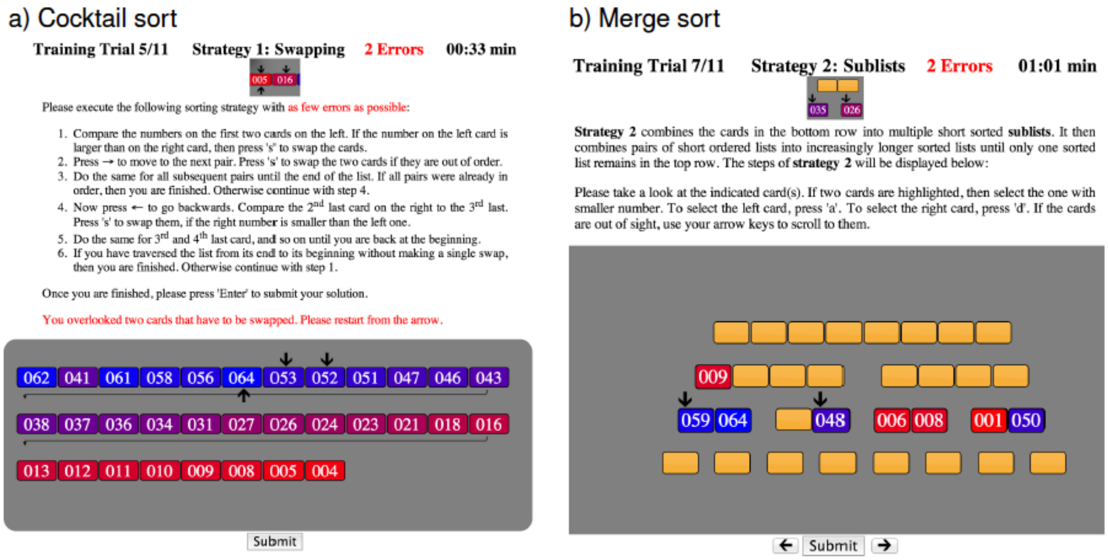
Pilot studies and simulations

To ensure that our experiment would be able to discriminate between rational metareasoning, SSL, RELACS, and SCADS, we simulated a number of candidate experiments. These simulations required a model of each strategy's performance. To obtain this execution time model, we conducted two pilot experiments in which we measured the execution time characteristics of cocktail sort (Pilot Experiment 1) and merge sort (Pilot Experiment 2) on different types of lists. The results of these pilot experiments will also allow us to determine when each strategy should be used to achieve optimal performance.

We recruited 200 participants on Amazon Mechanical Turk: 100 for each pilot experiment. Each participant was paid 75 cents for about 15 minutes of work. In Pilot Experiment 1 participants were required to follow the step-by-step instructions of the cocktail sort strategy (see Figure 2a). In Pilot Experiment 2 participants were required to

[2] A preliminary version of this study appeared in Lieder, Plunkett, et al. (2015).

follow the step-by-step instructions of the merge sort strategy (see Figure 2b). These two studies and all following experiments were conducted under the IRB protocol number 2015-05-755 entitled “Cognitive Research Using Amazon Mechanical Turk” approved by the institutional review board of the University of California, Berkeley. Participants were given detailed written instructions that precisely specified the strategy they had to execute. Furthermore, at each step the interface allowed only the correct next move of the required strategy and participants received feedback when they attempted an incorrect move. After completing four practice trials, participants were randomly assigned to sort a series of lists of varying lengths and presortedness. The lists were randomly generated so that each list was equally likely to be nearly sorted (1-20% inversions), unsorted (21-80% inversions), or nearly inversely sorted (81-100% inversions). Each list was equally likely to be very short (3-8 elements), short (9-16 elements), long (17-32 elements), or very long (33-56 elements). These lists were distributed across participants such that the total anticipated sorting time was between 10 and 20 minutes.



a) Cocktail sort

Training Trial 5/11 **Strategy 1: Swapping** **2 Errors** **00:33 min**

Please execute the following sorting strategy with as few errors as possible:

1. Compare the numbers on the first two cards on the left. If the number on the left card is larger than on the right card, then press 's' to swap the cards.
2. Press → to move to the next pair. Press 's' to swap the two cards if they are out of order.
3. Do the same for all subsequent pairs until the end of the list. If all pairs were already in order, then you are finished. Otherwise continue with step 4.
4. Now press ← to go backwards. Compare the 2nd last card on the right to the 3rd last. Press 's' to swap them, if the right number is smaller than the left one.
5. Do the same for 3rd and 4th last card, and so on until you are back at the beginning.
6. If you have traversed the list from its end to its beginning without making a single swap, then you are finished. Otherwise continue with step 1.

Once you are finished, please press 'Enter' to submit your solution.

You overlooked two cards that have to be swapped. Please restart from the arrow.

b) Merge sort

Training Trial 7/11 **Strategy 2: Sublists** **2 Errors** **01:01 min**

Strategy 2 combines the cards in the bottom row into multiple short sorted **sublists**. It then combines pairs of short ordered lists into increasingly longer sorted lists until only one sorted list remains in the top row. The steps of **strategy 2** will be displayed below:

Please take a look at the indicated card(s). If two cards are highlighted, then select the one with smaller number. To select the left card, press 'a'. To select the right card, press 'd'. If the cards are out of sight, use your arrow keys to scroll to them.

Figure 2: Interfaces used in Experiment 1 to train participants to perform (a) cocktail sort and (b) merge sort.

We used the measured sorting times to estimate how long comparisons and moves take for each strategy. For each list, we regressed the sorting times of each strategy on the number of comparisons and the number of moves that it performed on that list. The resulting model for the execution time T_{CS} of cocktail sort (CS) was

$$T_{\mathrm{CS}} = \hat{t}_{\mathrm{CS}} + \varepsilon_{\mathrm{CS}},$$
$$\hat{t}_{\mathrm{CS}} = 19.59 + 0.19 \cdot n_{\mathrm{comparisons}} + 0.31 \cdot n_{\mathrm{moves}} \quad (3),$$
$$\varepsilon_{\mathrm{CS}} \sim \mathcal{N}(0, 0.21 \cdot \hat{t}_{\mathrm{CS}}^2),$$

where $\hat{t}_{CS}$ is the expected execution time, ε_{CS} is the noise, $n_{\mathrm{comparisons}}$ is the number of comparisons and n_{moves} is the number of moves. For merge sort (MS) our data showed that both comparisons and moves took substantially longer:

$$T_{\mathrm{MS}} = \hat{t}_{\mathrm{MS}} + \varepsilon_{\mathrm{MS}},$$
$$\hat{t}_{\mathrm{MS}} = 13.98 + 1.10 \cdot n_{\mathrm{comparisons}} + 0.52 \cdot n_{moves} \quad (4)$$

$$\varepsilon_{\mathrm{MS}} \sim \mathcal{N}(0,0.15 \cdot \hat{t}_{MS}^2).$$

According to these execution time models (Equations 3-4) and the number of comparisons and moves required by these sorting strategies, people should choose merge sort for long and nearly inversely sorted lists and cocktail sort for lists that are either nearly-sorted or short. We will therefore classify people's strategy choices as adaptive when they conform to these rules and as non-adaptive when they violate them.

The execution time models of the two strategies also allowed us to simulate 104 candidate experiments according to rational metareasoning, SSL, RELACS, and SCADS. To apply SSL, RELACS, and SCADS to sorting strategies, we had to specify the reward function. We evaluated three notions of reward: i) correctness ($r \in \{-0.1, +0.1\}$[3], ii) correctness minus time cost ($r - \gamma \cdot\ t$, where t is the execution time and $\gamma = 1$ is the opportunity cost), and iii) reward rate ($^{r}/_{t}$). Each of the three theories (SSL, RELACS, and SCADS) was combined with each of these three notions of reward leading to 9 alternative models in total. Since the SCADS model presupposes that each problem is characterized by a collection of binary features we designed the following categories: short lists (length $\leq\ 16$), long lists (length $\geq\ 32$), nearly sorted lists (less than 10% inversions), and random lists (more than 25% inversions). According to the SCADS model, each problem can belong to multiple categories or none at all. To obtain an upper bound on how well the SCADS model can select sorting strategies, we also considered three SCADS models with categories that were optimized for this experiment. These categories were short-and-presorted, long-and-presorted, short-and-inverted, long-and-inverted, short-and-inverted, long-and-disordered, and short-and-disordered. Each of these categories is the conjunction of one attribute based on length (short means ≤ 25 and long means > 25) and one attribute based on presortedness (presorted means less than 25% inversions, inverted means more than 75% inversions, and disordered means 25—75% inversions). All associations between strategies and categories were initialized with a strength equivalent to one successful application, and the global strategy-reward associations were initialized in the same way. For consistency, the time cost parameter γ of the rational metareasoning model was also set to 1.[4]

Our simulations identified several candidate experiments for which rational metareasoning made qualitatively different predictions than SSL, RELACS, and SCADS. We selected the experimental design shown in Table 1 because it achieved the best tradeoff between discriminability and duration. For this experimental design, rational metareasoning predicted that the choices of more than 70% of our participants would demonstrate adaptive strategy selection, whereas the SSL, RELACS, and SCADS models all predicted that people would consistently fail to select their sorting strategy adaptively (see Figure 4).

[3] These specific values were taken from the SCADS model (Shrager & Siegler, 1998).

[4] The precise weighting of time cost versus error cost was irrelevant in these simulations because each sorting strategy was guaranteed to always generate a correct solution. Thus, there was no need to simulate how people estimate the time cost from experience.

Table 1: Design of Experiment 1.

Training Block

Trial Nr.	Strategy	Sequence Length	Inversions
1	Cocktail Sort	4	3
2	Cocktail Sort	8	7
3	Cocktail Sort	16	15
4	Cocktail Sort	16	1
5	Cocktail Sort	32	31
6	Merge Sort	4	3
7	Merge Sort	8	7
8	Merge Sort	16	15
9	Merge Sort	16	15
10	Cocktail Sort	32	1
11	Merge Sort	32	1

Choice Block

Trial Nr.	Sequence Length	Inversions
1	64	63
2	61	60
3	58	57
4	55	54
5	52	51
6	49	48
7	64	1
8	61	1
9	58	1
10	55	1
11	52	1
12	49	1
13	24	1
14	21	1
15	18	1
16	15	1
17	12	1

Methods

We recruited 100 participants on Amazon Mechanical Turk. Each participant was paid $1.25 for about 20 minutes of work. The experiment comprised three blocks: the training block, the choice block, and the execution block.

In the *training block*, each participant was taught the cocktail sort strategy and the merge sort strategy. In each of the 11 training trials summarized in

Table 1 participants were instructed which strategy to use. The interface shown in Figure 2 enforced that each of its step was executed correctly. Participants first practiced cocktail sort for five trials. Next, they practiced merge sort for four trials. These practice trials comprised nearly-reversely sorted lists of lengths 4, 8, and 16 and nearly-sorted lists of length 16 and 32. The nearly-sorted lists were created from ascending lists by inserting a randomly selected element at a random location. Nearly inversely sorted lists were created by applying the same procedure to a descending list. Finally, the last two trials contrasted the two strategies on two long, nearly-sorted lists (see

Table 1).

In the *choice block*, participants were shown 18 test lists and asked to choose which strategy (cocktail sort or merge sort) they would use if they had to sort it. To incentivize participants to choose the more efficient strategy, the instructions announced that in the following block one of their choices would be selected at random and they would have to execute it. The 18 test lists comprised six examples of each of three types of lists: long and nearly inversely sorted, long and nearly-sorted, and short and nearly-sorted (see Table 1). The order of these lists was randomized across participants.

In the *execution block*, one of the 12 short lists from the choice block was selected at random, and the participant had to sort it using the strategy they had selected for it.

Results

Our participants completed the experiment in 24.7 ± 6.7 minutes (mean ± standard deviation). In the training phase, the median number of errors per list was 2.45, and 95% of our participants made between 0.73 and 12.55 errors per list. The most important outcome was the relative frequency of adaptive strategy choices: On average, our participants chose merge sort for 4.9 of the 6 long and nearly inversely sorted lists for which it was optimal, that is 81.67% of the time. To quantify our uncertainty about this and subsequent frequency estimates we computed credible intervals based on a uniform prior (Edwards, Lindman, & Savage, 1963). According to this analysis, we can be 95% confident that the frequency with which people used merge sort on long nearly inversely sorted lists lies between 77.8% and 93.0%. By contrast, our participants chose merge sort for only 1.79 of the 6 *nearly-sorted* long lists for which it was inferior to cocktail sort (29.83% of the time, 95% credible interval: $[12.9\%, 32.4\%]$), and for only 1.62 of the 6 nearly-sorted short lists for which it was also inferior (27.00% of the time, 95% credible interval: $[16.7\%, 40.4\%]$); see Figure 3A. Thus, our participants chose merge sort significantly more often than cocktail sort when it was superior ($p < 10^{-10}$; Cohen's $w = 6.12$). But, when merge sort was inferior, they chose it significantly less often than cocktail sort ($p < 10^{-7}$, Cohen's $w = 6.33$). Overall, 83% of our participants chose merge sort more often when it was the superior strategy than when cocktail sort was the superior strategy and vice versa (95% credible interval: $[74.9\%; 89.4\%]$; see Figure 3). The high frequency of this adaptive strategy choice pattern provides strong evidence for the hypothesis that people's strategy choices are informed by features of the problem to be solved, because it would be extremely unlikely otherwise ($p < 10^{-11}$, Cohen's $w = 6.60$). This finding was predicted by our rational metareasoning model of strategy selection which achieved adaptive strategy selection in 70.5% of the simulations ($p < 10^{-14}$) and the SCADS model with optimized categories and the VOC-based reward function (performance minus time cost) which achieved adaptive strategy selection in 59.0% of the simulations ($p < 10^{-5}$) but not by any of the other SCADS, RELACS, and SSL models (all $p \geq 0.17$). Figure 3A compares the proportion of participants who chose their sorting strategy adaptively with the models' predictions. The non-overlapping credible intervals suggest that we can be at least 95% confident that people's strategy choices were more adaptive than predicted by SSL, RELACS, or SCADS and a series of t-tests confirmed this interpretation (all $p < 0.001$). While the frequency of adaptive strategy choices predicted by rational metareasoning ($70.5 \pm 3.2\%$) was also significantly higher than for any of the alternative models (all $p < 0.01$), our participants chose their strategies even more adaptively than our rational metareasoning model predicted (83.0% vs. 70.5%, $t(298) = 2.34$, $p = 0.01$). Like people, rational metareasoning selected merge sort for significantly more than half of the lists that were long and inverted ($p < 10^{-6}$) but for significantly less than half of the lists that were long and presorted ($p < 10^{-15}$) or short and presorted ($p < 10^{-15}$). As shown in Figure 3B, none of the alternative models captured this pattern.

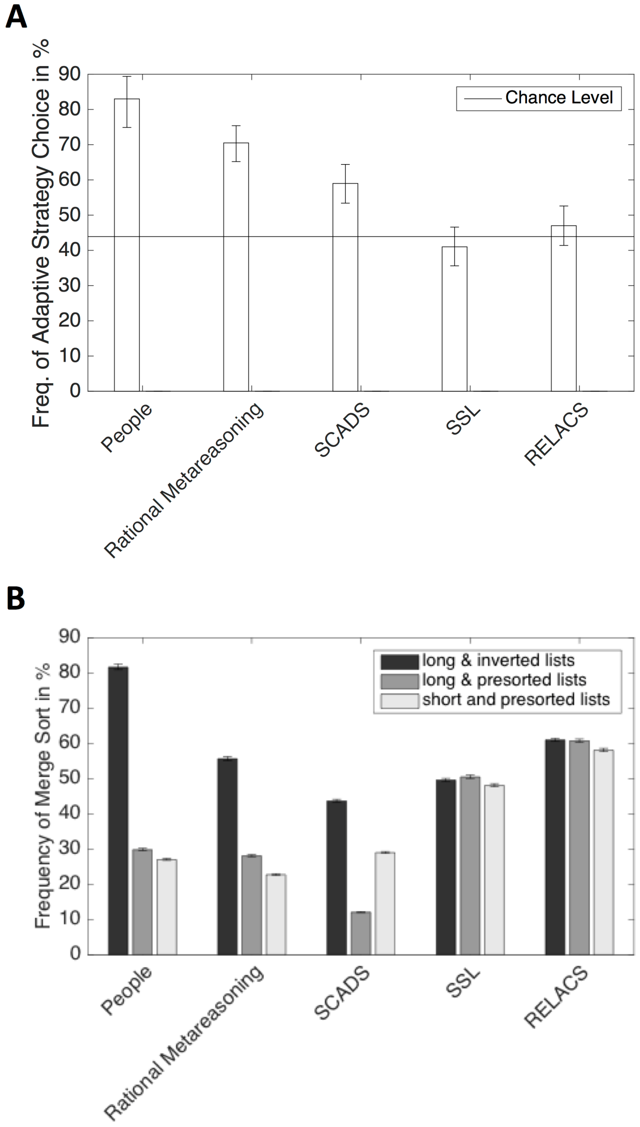


Figure 3: Pattern of strategy choices in Experiment 1. A: Percentage of participants who chose merge sort more often when it was superior than when it was not. Error bars indicate 95% credible intervals. The results for SCADS, SSL, and RELACS correspond to the version of the respective model that achieved the highest frequency of adaptive strategy selection. B: Relative frequency with which humans and models chose merge sort by list type.

Our model has four components: i) choosing strategies based on their VOC by trading off expected performance versus expected cost, ii) learning to predict the performance of strategies from features of individual problems, iii) learning separate predictive models of computational effort and reward, and iv) meta-cognitive exploration by Thompson sampling. To determine which components of our model are critical to its ability to choose strategies adaptively, we created additional models by removing each of the four components in turn. This resulted in five additional models: one rational metareasoning model without features, one rational metareasoning model without exploration, two models that choose strategies based on criteria other than the VOC, and one model that approximated the VOC directly without learning to predict execution time and reward separately. The last three models use the same reward functions as the three instantiations of each of the previous theories of strategy selection learning: reward only, reward rate, and reward minus time cost; while the first two models choose strategies based on a criterion other than the VOC, the last model learns a model-free approximation to the VOC without learning to predict either deliberation time or accuracy.

To evaluate these "lesioned" models, we simulated the sorting experiment according to each of them and measured how often the resulting strategy choices were adaptive (see Supplementary Figure 4 in the Online Supplementary Material). We found that the features and the VOC-based strategy selection mechanism were necessary to capture human performance. Exploration and learning separate predictive models for execution time and accuracy were not necessary to capture human performance in the sorting task, but they were necessary to capture human performance in the experiments simulated below; detailed statistical analyses are provided in the Online Supplementary Material.

Discussion

We evaluated rational metareasoning against three existing theories of human strategy selection. We found that the predictions of rational metareasoning were qualitatively correct and that its choices came close to human performance. By contrast, the nine alternative models instantiating previous theories completely failed to predict people's adaptive strategy choices in our experiment: The RELACS and SSL models do not represent problem features and thus cannot account for people's ability to learn how those features affect each strategy's performance. The basic associative learning mechanism assumed by SSL and RELACS was maladaptive in Experiment 1 because cocktail sort was faster for most training lists but substantially slower for the long, nearly inversely sorted test lists.

The primary advantage allowing our model to perform better than SSL and RELACS is that it leverages problem features that distinguish the lists for which cocktail sort is superior from the lists for which merge sort is superior. If SSL and RELACS were applied two either set of lists separately, they would learn to identify the correct strategy for each of them. However, in the real world, problems rarely come with a single label that identifies the correct strategy. Instead, the correct strategy has to be inferred from a combination of multiple features (e.g., length and presortedness) none of which is sufficient to choose correct strategy on its own. Our rational metareasoning model handles this challenge but SSL and RELACS do not address it yet.

The SCADS model failed mainly because its associative learning mechanism was fooled by the imbalance between the training examples for cocktail sort and merge sort. Furthermore, the strategy selection component of the SCADS model can neither extrapolate nor capture the non-additive interaction between length and presortedness.

Our findings suggest that people leverage the features of individual problems to learn how to select strategies adaptively. The success of the rational metareasoning model and its evaluation against lesioned metareasoning models suggests that our hypothesis that people learn to predict the VOC of alternative strategies from the features of individual problems may be able to account for the adaptive flexibility of human strategy selection.

In contrast to the sorting strategies in Experiment 1, most cognitive strategies operate on internal representations. In principle, strategies that operate on internal representations could be selected by a different mechanism than strategies that operate on external representations. However, there are two reasons to expect our conclusions to transfer: First, people routinely apply strategies that they have applied to external objects to their internal representations of those objects. For instance, mental arithmetic is based on calculating with fingers. Thus, the strategies people use to order things mentally might also be based on the strategies they use to sort physical objects. Second, strategy selection can be seen as an instance of metacognitive control, and metacognitive processes tend to be domain general. In the following sections, we show that our conclusions do indeed transfer to cognitive strategies that operate on internal representations.

Cognitive flexibility in complex decision environments

People are known to use a wide repertoire of different heuristics to make decisions under risk (Payne, Bettman, & Johnson, 1993). These strategies include *fast-and-frugal* heuristics which, as the name suggests, perform very few computations and use only a small subset of the available information (Gigerenzer & Gaissmaier, 2011). For instance, the lexicographic heuristic (LEX) focuses exclusively on the most probable outcome that distinguishes between the available options and ignores all other possible outcomes. Another fast-and-frugal heuristic that people might sometimes use is Elimination-By-Aspects (EBA; Tversky, 1972). Here, we used the deterministic version of EBA described by Payne et al. (1988). This heuristic starts by eliminating options whose payoff for the most probable outcome falls below a certain threshold. If more than one option remains, EBA repeats the elimination process with the second most probable outcome. This process repeats until only one option remains or all outcomes have been processed. After the elimination step EBA chooses one of the remaining outcomes at random. In addition to fast-and-frugal heuristics, people's repertoire also includes more time consuming but potentially more accurate strategies such as the weighted-additive strategy (WADD). WADD first computes each option's expected value, and then chooses the option whose expected value is highest.

In addition to gradually adapting their strategy choices to the structure of the environment (Rieskamp & Otto, 2007) people can also flexibly switch their strategy as soon as a different problem is presented. Payne et al. (1988) provided a compelling demonstration of this phenomenon in risky choice: Participants chose between multiple gambles described by their possible payoffs and their respective probabilities. There was a fixed set of possible outcomes that occurred with known probabilities and the gambles differed in the payoffs they assigned to these outcomes. Participants were presented with four types of decision problems that were defined by the presence or absence of a time constraint (15 seconds vs. none) and the dispersion of the outcomes' probabilities (low vs. high); high dispersion means that some outcomes are much more probable than others, whereas low dispersion means that all outcomes are almost equally likely. Ten instances of each of the four problem types were intermixed in random order. The outcomes' payoffs ranged from $0 to $9.99, and their values and probabilities were stated numerically. Payne et al. (1988) used process tracing to infer which strategies their participants were using: The payoffs and their probabilities were revealed only when the participant clicked on the corresponding cell of the payoff matrix displayed on the screen, and all mouse clicks were recorded. This allowed Payne and colleagues to measure how often people used the fast-and-frugal heuristics (LEX and EBA) for different types of problems by the percentage of time people spent on the options' payoffs for the most probable outcome. For the expected-value strategy WADD this proportion is only 25%, but for the fast-and-frugal heuristics LEX and EBA it can be up to 100%. The experiment revealed that people adaptively switch decision strategies in the absence of feedback: When the dispersion of outcome probabilities was high, people focused more on the most probable outcome than when all outcomes were almost equally probable. Time pressure also increased people's propensity for such selective and attribute-based processing; see Figure 4. Thus, participants appeared to use fast-and-frugal heuristics more frequently when they had to be fast and when all but one or two outcomes were extremely improbable. This makes sense because the fast-and-frugal heuristics LEX and EBA are fast precisely because they focus on the most predictive attributes instead of integrating all attributes.

We investigated whether rational metareasoning can account for people’s adaptive flexibility in this experiment. To do so, we simulated the experiment by applying our model to the selection between the ten decision strategies considered by Payne et al. (1988) including WADD and fast-and-frugal heuristics such as LEX and EBA. To simulate each strategy’s execution time we counted how many elementary operations (Johnson & Payne, 1985) it would perform on a given problem and assumed that each of them takes one second. This allowed us to simulate the effect of the time limit on a strategy’s performance by having each strategy return its current best guess when it exceeds the time limit (Payne et al., 1988). For the purpose of strategy selection learning, our model represented each decision problem by five simple and easily computed features: the number of possible outcomes, the number of options, the number of inputs per available computation, the highest outcome probability, and the difference between the highest and the lowest payoff. Our model used these features to learn a predictive model of each strategy's relative reward

$$r_{\mathrm{rel}}(s;o) = \frac{V(s(D),o)}{\max_{a} V(a,o)},$$

where $s(D)$ is the gamble that strategy s chooses in decision problem D, $V(c,o)$ is the payoff of choice c if the outcome is o, and the denominator is the highest payoff the agent could have achieved given that the outcome was o. To choose a strategy, the predicted relative reward $\hat{r}_{\mathrm{rel}}$ is translated into the predicted absolute reward $\hat{r}$ by the transformation

$$\hat{r} = \min\{r_{min} + (r_{max} - r_{min}) \cdot \hat{r}_{rel}, r_{max}\},$$

where r_{min} and r_{max} are the smallest and the largest possible payoff of the current gamble respectively. The model then integrates the predicted absolute reward and the predicted time cost into a prediction of the strategy’s VOC according to Equation 2 and chooses the strategy with the highest VOC as usual. The priors on all feature weights of the score and execution time models were standard normal distributions. The simulation assumed that people knew their opportunity cost and did not have to learn it from experience. Rather than requiring the model to learn the time cost as outlined above, the opportunity cost was set to \$7 per hour and normalized by the maximum payoff (\$10) to make it commensurable with the normalized rewards.

To compare people's strategy choices to rational metareasoning, we performed 1000 simulations of people's strategy choices in this experiment. In each simulation, we modeled people's prior knowledge about risky choice strategies by letting our model learn from ten randomly generated instances of each of the 144 types of decision problems considered by Payne et al. (1988). We then applied rational metareasoning with the learned model of the strategies' execution time and expected reward to a simulation of Experiment 1 from Payne et al. (1988). On each simulated trial, we randomly picked one of the four instances and generated the payoffs and outcome probabilities according to the problem type: Outcome distributions with low dispersion were generated by sampling outcome probabilities independently from the standard uniform distribution and dividing them by their sum. Outcome distributions with high dispersion were generated by sampling the outcome probabilities sequentially such that the second largest probability was at most 25% of the largest one, the third largest probability was at most 25% of the second largest one, and so on. Since the participants in this experiment received no feedback, our simulation assumed no learning during the experiment.

To evaluate our theory against alternative hypotheses, we also ran 1000 simulations according to SCADS. To evaluate our theory against alternative hypotheses, we also ran 1000 simulations according to the SCADS model. We did not evaluate SSL or RELACS because these theories do not distinguish different kinds of problems and hence cannot account for the phenomena observed by Payne et al. (1988).

The SCADS model was equipped with nine categories (time pressure, no time pressure, many options (> 3), few options (≤ 3), many possible outcomes (> 3), few possible outcomes (≤ 3), high stakes (range of payoffs ≥ 50% of highest payoff), low-stakes (range of payoffs $\leq$ 10% of highest payoff), and non-compensatory (largest outcome probability > 0.5)). As before, we considered three instances of SCADS whose reward functions were either the relative payoff, the relative payoff minus the opportunity cost of the strategy's execution time, or the reward rate. The SCADS model's category-specific and global strategy-reward associations were initialized with strengths equivalent to one observation per strategy.

We found that rational metareasoning correctly predicted that time-pressure and probability dispersion increase people's propensity to use the fast-and-frugal heuristics LEX and EBA; see Figure 4. Time pressure increased the predicted frequency of fast, attribute-based processing by 29.69% ($t(1998) = 9.70, p < 10^{-15}$), and high dispersion of the outcome probabilities increased the predicted frequency of fast, attribute-based processing by 44.11% ($t(1998) = 14.41, p < 10^{-15}$). Furthermore, their strategy choices only change in response to reward or punishment but the experiment provided neither. The SCADS model can choose strategies adaptively in principle, but in our simulations its strategy choices were dominated by the global, problem-independent associations. Consequently, our SCADS models always converged onto a single strategy during the training phase and continued to do so in the test trials. Hence, the predicted effects of time pressure (-0.1 to 0%, all $p \geq 0.4955$) and dispersion (0% to 0.05%, all $p \geq 0.4978$) were not significantly different from zero. In conclusion, rational metareasoning can account for adaptive flexibility in decision-making under risk but SSL, RELACS, and SCADS cannot. These results suggest that rational metareasoning can capture the adaptive flexibility of people's strategy choices not only for behavioral strategies that manipulate external representations but also for cognitive strategies that operate on internal representations.

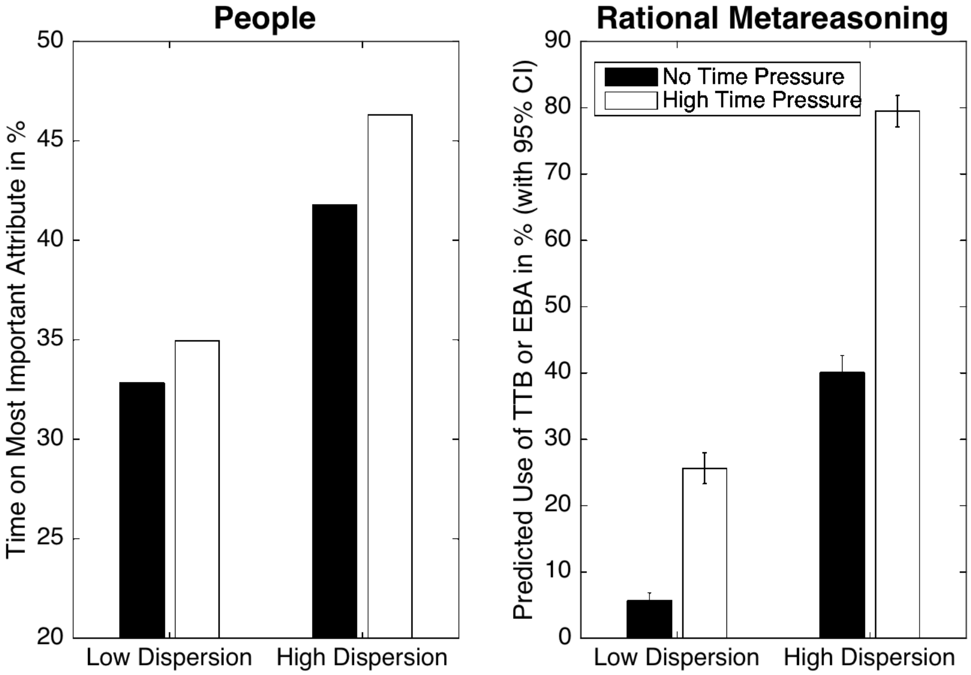


Figure 4: Rational metareasoning predicts the increase in selective attribute-based processing with dispersion and time pressure observed by Payne et al. (1988).

To evaluate which components of rational metareasoning are critical to capture people's adaptive decision-making, we lesioned our model by separately removing each of its four components. We found that the feature-based problem representations and exploration were critical to the model's adaptive strategy choices but learning separate models of the costs and benefits and choosing strategies based on the VOC was not; for more detail see Appendix B. Although learning about the time cost was not necessary to perform well in the experiment by Payne et al. (1988), there are other scenarios, such as the sorting experiment, where this is critical.

Rational strategy selection is learned from experience

In the previous sections, we have shown that our feature-based strategy selection model can explain people's ability to choose cognitive and behavioral strategies to flexibly adapt how they process information to the requirements of their current situation. According to our theory, people acquire this ability by learning an internal predictive model of each strategy's performance. In this process, people should gradually learn to perform more valuable computations and fewer computations whose costs outweigh their benefits. In other words, people should learn to make increasingly more rational use of their finite time and computational resources. This hypothesis makes four predictions:

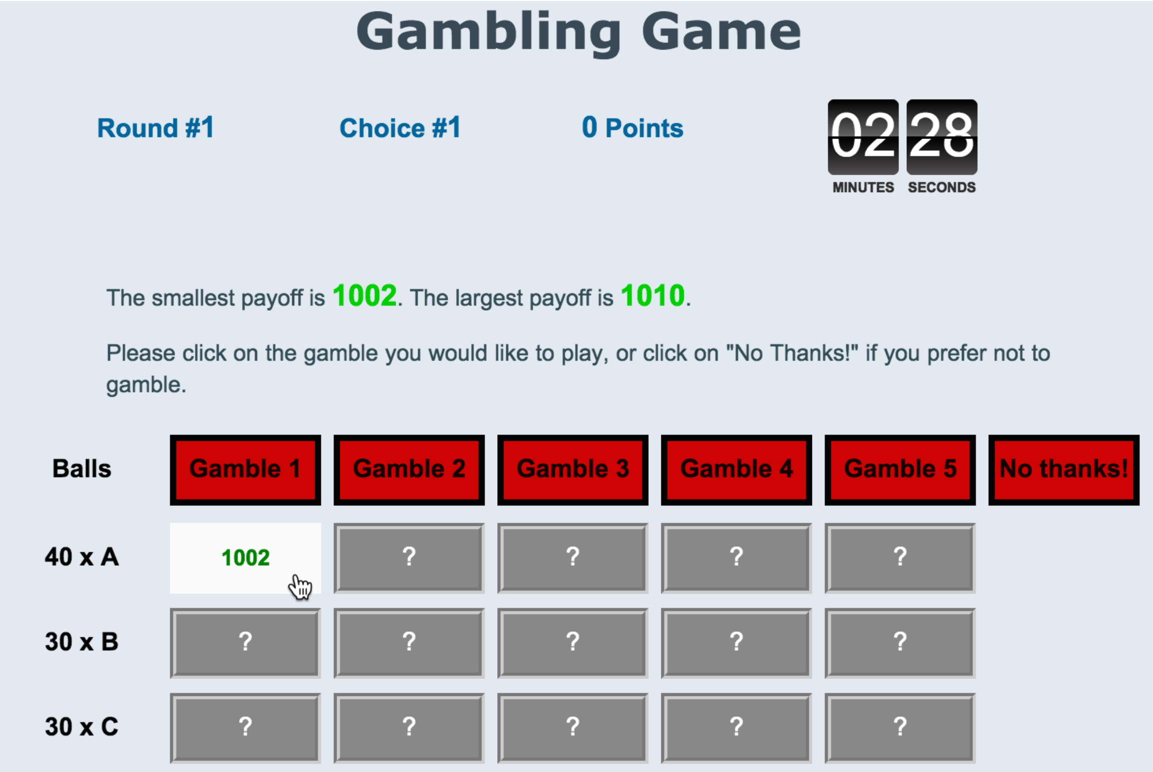
1. People learn to perform fewer computations whose time cost outweighs the resulting gain in decision quality.
2. People learn to perform more computations whose expected gain in decision quality outweighs their time cost.
3. Ecological rationality increases with learning: people gradually learn to adapt their strategy choices to the structure of their environment.
4. Adaptive flexibility increases with learning: people learn to use different strategies for different kinds of problems.

We test these predictions in the remainder of this section.

Experiment 2: When people think too much they learn to think less

The goal of Experiment 2 was to test our model's prediction that people will learn to deliberate less and decide more quickly when they are placed in an environment where the cost of deliberation outweighs its benefits.

Methods. We recruited 100 adult participants on Amazon Mechanical Turk. Participants were paid $0.75 for 15 minutes of work and could earn a bonus of up to $2 for their performance on the task; the average bonus was $1.15 and its standard deviation was $0.73. The experiment was structured into three blocks: a pretest block, a training block, and a posttest block. Participants received feedback about the outcomes of their choices in the training block but not in pretest or the posttest block. Each block lasted four minutes, and the participants' task was to win as many points as they could.



Gambling Game

Round #1 Choice #1 0 Points

02 28
MINUTES SECONDS

The smallest payoff is 1002. The largest payoff is 1010.

Please click on the gamble you would like to play, or click on "No Thanks!" if you prefer not to gamble.

Balls	Gamble 1	Gamble 2	Gamble 3	Gamble 4	Gamble 5	No thanks!
40 x A	1002	?	?	?	?	
30 x B	?	?	?	?	?	
30 x C	?	?	?	?	?	

Figure 5: Screenshot of example trial in the pretest phase of Experiment 2.

Figure 5 shows a screenshot of an example trial in the pretest phase. In each trial, participants were shown a number of gambles. They could either choose one of the gambles or skip the decision and move on to the next trial without receiving a payoff. As soon as the participant responded the next trial was shown. The number of trials was solely determined by how quickly the participant responded on each trial. On each trial, the decision problem was equally likely to belong to either of the four types summarized in Table 2. The four problem types differed primarily in the range of possible payoffs (low stakes, vs. high stakes, vs. all positive, vs. all negative), and on each trial this range was shown as a cue (see Figure 5). Critically, as shown in Table 2, the problem types and their frequencies were chosen such that the best approach was to skip trials where all outcomes were negative, choose randomly on trials where all outcomes were positive, and minimize the time spent on the high-stakes and the low-stakes problems by choosing randomly or skipping them altogether.

Table 2

Frequency and properties of the four types of decision problems used in Experiment 2.

Problem Type	Frequency	Worst Outcome	Best Outcome	Optimal Strategy
All great	25%	990	1010	random choice
All bad	25%	-1010	-1000	Disengagement
High Stakes	25%	-1000	1000	Disengagement
Low Stakes	25%	-10	10	Disengagement

Note: All gambles were compensatory.

The number of outcomes was 3, 4, or 5 with probability 0.25, 0.50, and 0.25 respectively. The number of gambles was either 4 or 5 with equal probability. Given the number of outcomes and gambles, the payoffs were sampled uniformly from the problem type's range of payoffs given in Table 2. The outcome probabilities were sampled independently from the payoffs. Concretely, if there were k outcomes, then the first $k-1$ outcome probabilities were sampled by a stick-breaking process where the relative length of each new stick was sampled from a uniform distribution. The probability of the k-th outcome was set to 1 minus the sum of the first $k-1$ probabilities.

Model Predictions. To simulate people's choice of decision strategies and how it changes with learning, we combined our rational process model of strategy selection learning with the 10 decision strategies considered by Payne et al. (1988): the weighted-additive strategy, the equal weight strategy, satisficing, choosing at random, the majority of confirming dimensions strategy, the lexicographic heuristic (take-the-best), the semi-lexicographic heuristic, elimination-by-aspects, as well as two hybrid strategies that combine elimination-by-aspects with the weighted-additive strategy and the majority of confirming dimensions strategy respectively. Two additional strategies allowed the decision-maker to choose at random and skip the trial without deliberation respectively. The model's prior on the reward rate was a normal distribution with a mean of 1 point per second and a precision equivalent to 1 minute's worth of experience in the task. The priors on the regression coefficients and the error variance of the agent's predictive model of the strategies' performance were the same as in the simulations of the experiment by Payne et al. (1988). The features of the agent's predictive model combined those used to simulate the experiment by Payne et al. (1988) with four indicator variables signaling the presence or absence of the cues associated with the four types of gambles. Using these parameters, we ran 200 simulations of the experiment according to each model.

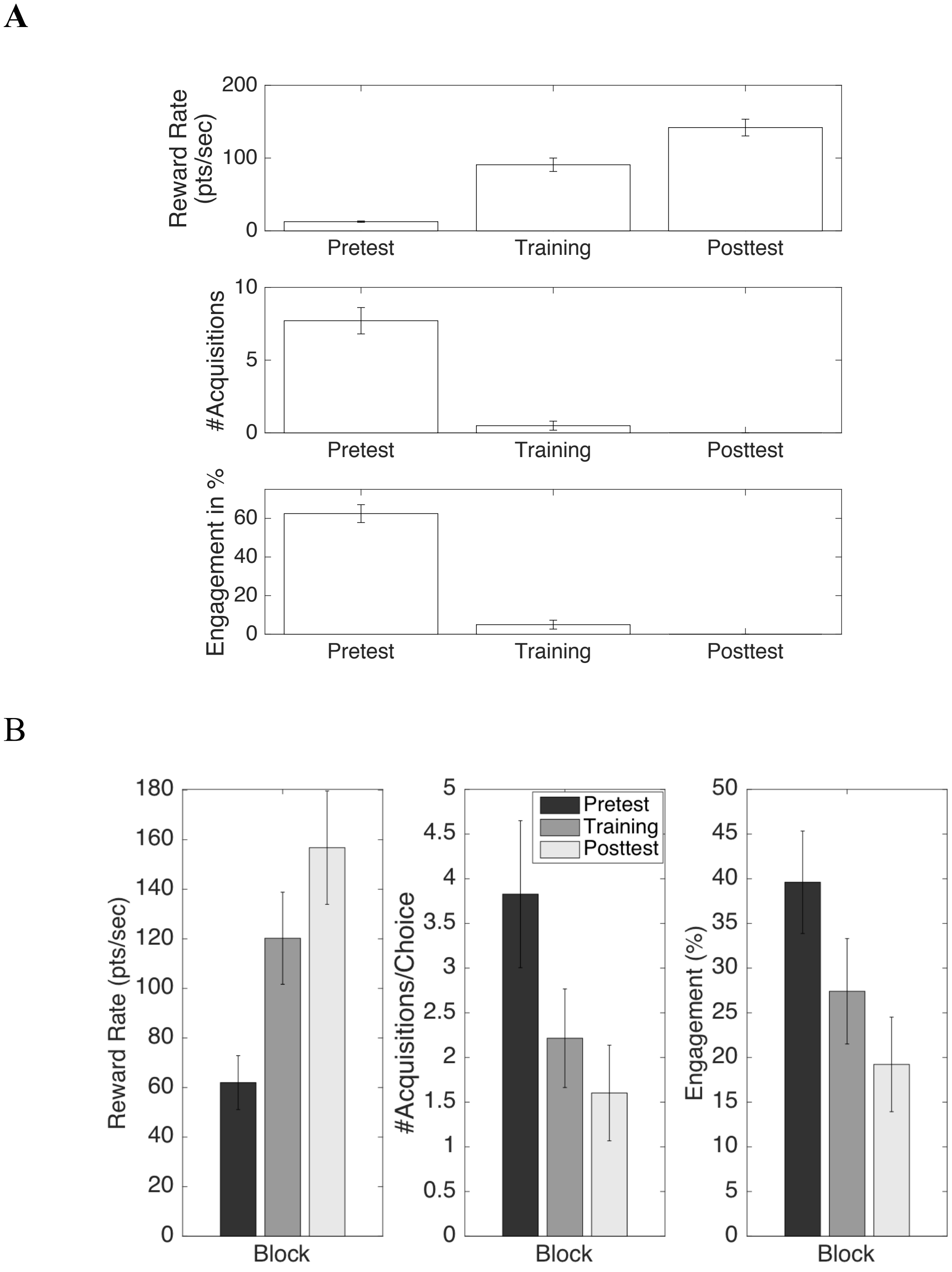


Figure 6: Experiment 2: Learning when not to engage in effortful decision-making. A: Predictions of rational metareasoning for Experiment 2. B: The empirical findings of Experiment 2 confirmed the three qualitative model predictions.

As shown in Figure 6A, our rational model predicted that participants should learn to decide more quickly and thereby win increasingly more points per second by engaging in deliberation less often and acquiring fewer pieces of information. Since the simulated decision-maker estimates its reward rate by Bayesian inference as defined above, it gradually realizes that its opportunity cost is very high. In addition, the simulated decision-maker learns that deliberate strategies are slow, and that the random strategy performs about as well as deliberation when all outcomes are similar. Hence, the simulated decision-maker eventually learns to avoid deliberating, to skip problems with negative payoffs, and to apply the random strategy when all outcomes are great.

Results. To test our hypothesis that people learn to deliberate less, we classified the participants' response patterns into three categories: The response strategy on a trial was categorized as *random choice* if the participant chose one of the gambles without inspecting any of the outcomes. If the participant chose "No thanks!" without inspecting the outcomes, then the response strategy was classified as *disengaged*. Finally, if the participant clicked on at least one of the outcome boxes, then the response was categorized as *engaged*. We measured our participants' performance on the task by three metrics: *engagement*, *reward rate,* and *adaptive randomness*. Engagement was defined as the proportion of trials on which participants were engaged; the reward rate is the number of points earned per second; and adaptive randomness was measured by the frequency of random choice in problem type 1 (*all great*) minus the frequency of random choice on problems of types 2 (*all bad*) and 3 (*high stakes*); see Table 2. Our model predicted that participants' reward rate and adaptive randomness would increase significantly from the pretest to the posttest while their engagement decreases.

As shown in Figure 6B, we found that the learning induced changes in our participants' strategy choices were consistent with our theory's predictions. There was a significant increase in the participants' average reward rate ($t(99) = 9.98, p < 10^{-15}$; Cohen's $d = 1.00$) as they learned to process less information ($t(99) = -4.80, p < 10^{-5}$; Cohen's $d = -0.48$) and their engagement decreased significantly ($t(98) = -7.89, p < 10^{-11}$; Cohen's $d = -0.79$). Even though participants acquired increasingly less information, their average reward per decision did not change significantly from the first block to the last block ($t(98) = 0.69, p = 0.49$; Cohen's $d = 0.07$).

To examine whether the effect of learning on the number of computations performed by our participants depended on the problem type we ran a 2x2 mixed-effects, repeated-measures ANOVA with the average number of information acquisitions for a given problem type in a given block as the dependent variable and the problem type and the block number as independent variables. The main effect of the problem type was significant ($F(3,1184) = 23.01, p < 10^{-13}$) suggesting that participants' information acquisition strategies differed significantly between the four types of decision problems (see Figure 7A): In high-stakes decisions, participants inspected 2.95 ± 0.55 outcomes on average, but on the trials where all outcomes were equally bad they inspected only about 0.5 potential payoffs (Cohen's $d = 1.96$). For low-stakes decisions and decisions in which all possible outcomes were great participants inspected an intermediate number of outcomes (about 1.5 inspected outcomes on average, Cohen's $d = 1.21$ and $d = 1.18$

respectively). The number of information acquisitions changed significantly across the three blocks of the experiment ($F(1{,}1184) = 23.64, p < 10^{-5}$). Concretely, information acquisition decreased by 1.4 pieces of information per block ($t(1184) = -4.86, p < 10^{-6}$; Cohen's $d = -0.14$). There was a statistically significant interaction between problem type and block number ($F(3{,}1184) = 4.74, p = 0.003$) indicating that the number of information acquisitions decreased more strongly for some problem types than for others. This decrease was statistically significant for problems in which all outcomes are great ($t(99) = -3.30, p < 0.001$, Cohen's $d = -0.33$), problems in which all outcomes are bad ($t(99) = -5.15,\ p < 10^{-6}$, Cohen's $d = -0.52$), and the high-stakes decision problems ($t(99) = -5.06, p < 10^{-6}$, Cohen's $d = -0.51$). But for the low-stakes problems the decrease was weaker and not statistically significant ($t(99) = -1.50, p = 0.07$, Cohen's $d = -0.15$).

The observed decrease in the number of information acquisitions was partly driven by a decrease in the frequency with which people engaged with the decision problems by inspecting at least one of their payoffs. As shown in Figure 7A, the proportion of decision problems in which people inspected at least one of the payoffs dropped from 37% in the pretest to 19% in the posttest. To test whether learning decreased the number of computations that people perform above and beyond the effect of disengagement, we repeated the analysis of variance described above for only those trials on which people engaged with the decision problem (see Figure 7B). We found that the main effect of the block number was still highly significant ($F(1{,}659) = 8.08, p = 0.005$). The estimated decrease in information acquisition on trials on which people engaged with the decision problem was 1.1 pieces of information per block (95% CI: $[-1.80\,, -0.33]$, $t(659) = -2.84, p = 0.005$, Cohen's $d = -0.11$) and this value was not significantly different from the average decrease across all trials (1.4 acquisitions/block, 95% CI: $[-1.60, -0.68]$). There was also a significant interaction between the block number and problem type ($F(3{,}659) = 2.61, p = 0.05$).

Furthermore, we found a significant increase in adaptive randomness ($t(97) = 7.21, p < 10^{-10}$, Cohen's $d = 0.73$). This means that our participants learned to selectively apply the random choice strategy to the *all great* problems (see Figure 7C). Consistent with this finding, the frequency of random choice increased on the *all great* trials ($t(97) = 6.61, p < 10^{-8}$, Cohen's $d = 0.67$) but decreased on all other trial types ($t(98) = -2.77, p = 0.003$, Cohen's $d = -0.28$).

Finally, we investigated whether people learn to prioritize the most probable outcome over less probable outcomes. To do so, we recorded the rank of the probability of the outcome participants inspected first and averaged it by block. The rank of the most probable outcome is one, the rank of the second most probable outcome is two, etc. On average, people inspected the second most probable outcome first. This is consistent with the interpretation that our participants sometimes used strategies that prioritize the most probable outcomes and sometimes used strategies that do not. There was a very small and almost statistically significant decrease in the rank of the probability of the outcome inspected first from 2.33 ± 0.05 in the pretest to 2.15 ± 0.08 in the posttest ($t(59) = -1.67,\ p = 0.05$; Cohen's $d = 0.22$).

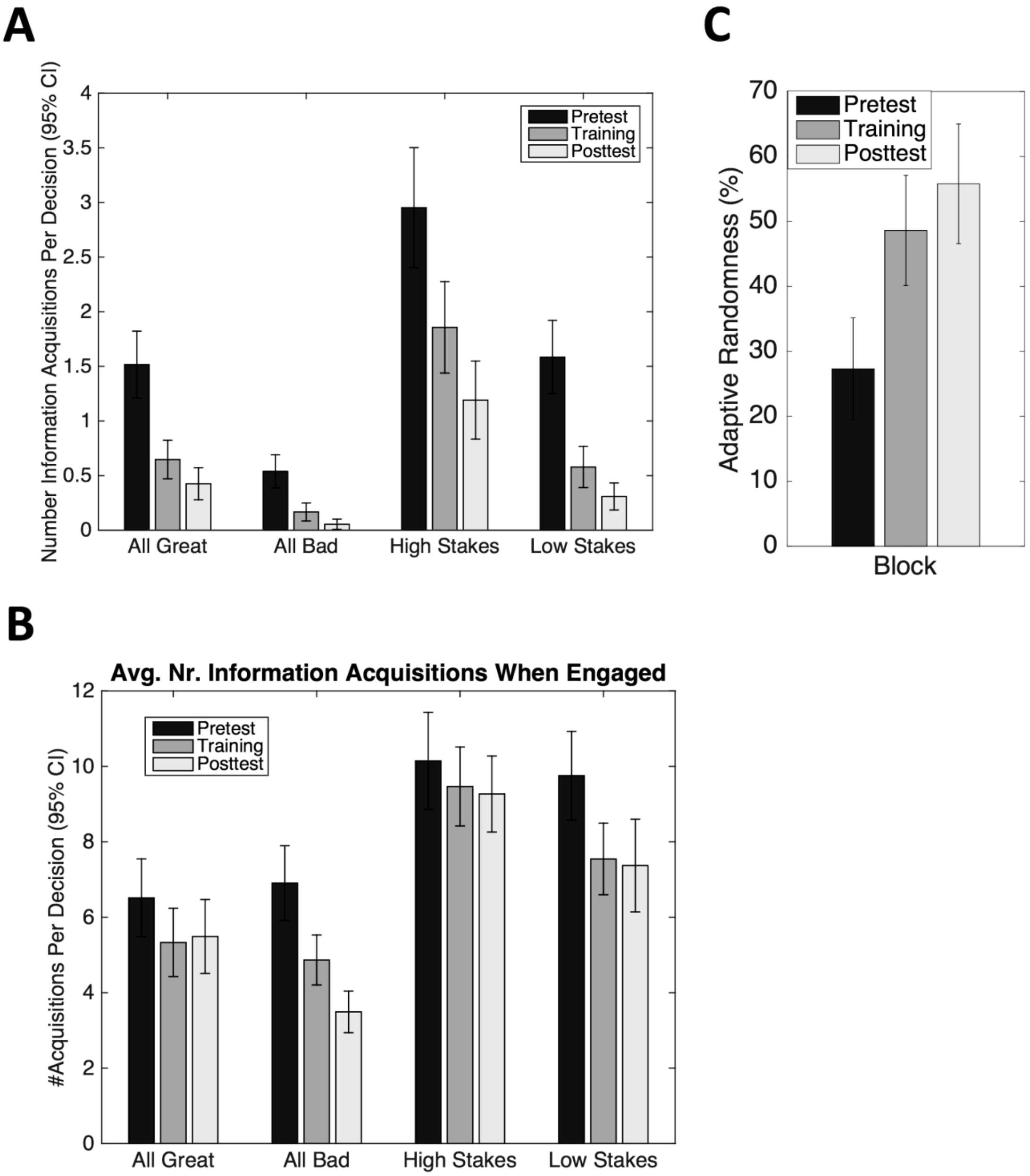


Figure 7: Adaptive disengagement in Experiment 2. A: Average number of information acquisitions by block and problem type. B: Number of information acquisitions when engaged. C: Adaptive randomness increased as participants learned to apply the random choice strategy more often to problems where all outcomes were great and less often to other problems.

In summary, Experiment 2 placed participants in an environment where maximizing the reward rate required choosing without deliberation, and the participants learned to reap increasingly higher reward rates by acquiring increasingly fewer pieces of information, choosing at random when all outcomes were great and to skipping all other problems. There was also a trend towards learning to prioritize the most probable outcome. All of these effects are consistent with the hypothesis that people learn to make increasingly more rational use of their finite time and computational resources.

Model Comparisons. While our findings were qualitatively consistent with the model predictions there were quantitative differences: People tended to outperform the model in terms of the reward rate in the pretest block, and their average number of acquisitions and frequency of engaging in deliberation changed less than predicted by rational metareasoning (compare Figure 6A vs. Figure 6B, and see the Supplementary Online Material for a more detailed comparison).

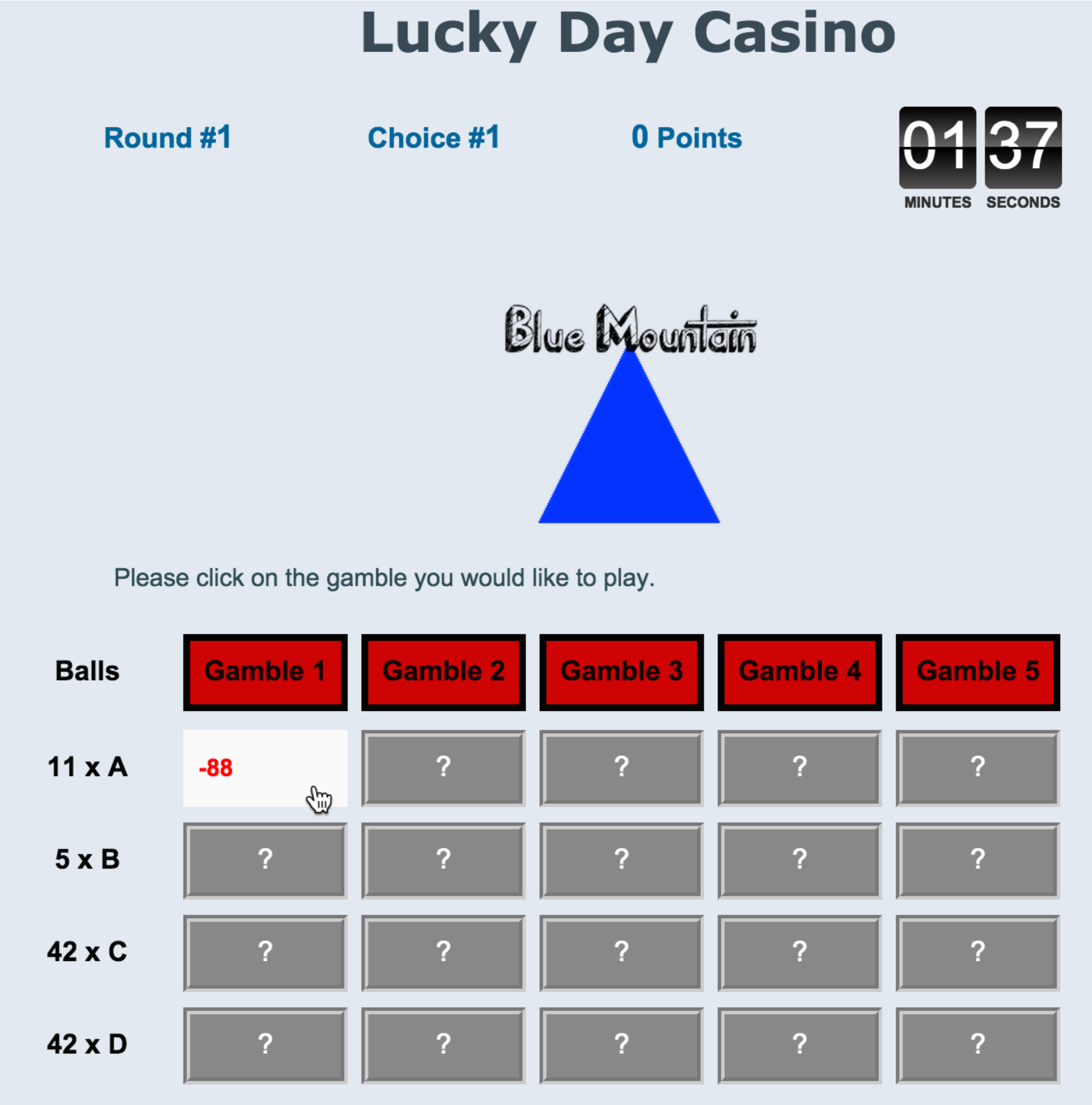
To evaluate our rational metareasoning model against the 14 alternative models described above, we ran 200 simulations of Experiment 2 according to each of the models. For each model, we performed six one-sample t-tests to determine whether it captured the increase in reward rate, the decrease in the number of acquisitions, and the decrease in the frequency of engagement from block 1 to block 2 and from block 2 to block 3, and one t-test to evaluate whether the model captured that people acquired more pieces of information on high-stakes problems than on other kinds of problems. We found that while our rational metareasoning model captured all of these effects, none of the SCADS, RELACS, or SSL models were able to capture all four effects simultaneously. The only component of the metareasoning model that was not necessary to capture human performance in Experiment 2 were the features. The reason why the lesioned metareasoning model without features could perform well is that the explicitly stated payoff ranges were sufficient for choosing strategies adaptively. Critically, none of the other lesioned metareasoning models were able to capture human performance. This suggests that all other components of our rational metareasoning model—choosing strategies based on the VOC, exploration, and learning separate predictive models of execution time and reward—are necessary to capture people's ability to adapt to the decision environment of Experiment 2. For a more detailed summary of these simulation results, please see Appendix B.

Discussion. The observation that sometimes people are cognitive misers poses a challenge to most rational models, but our model predicted it correctly. According to our model, people become faster and less accurate at a challenging task when the difference between the rewards for good versus bad performance is small compared to how much time it would take to perform better. The observation that over time participants came to engage less with all four types of problems could also be interpreted as a general disengagement from the experiment rather than a rational adaptation to the structure of the decision environment. To disambiguate rational adaptation from disengagement we designed an additional experiment in which our theory predicts that people should learn to invest increasingly more time and effort.

Experiment 3: Learning to deliberate more

The goal of Experiment 3 was to test our model's prediction that people learn to deliberate more when they initially think too little. To create a situation where people think too little, we first put them in an environment whose reward rate was so high that deliberating on low-stakes problems was a waste of time and then changed the environment so that low-stakes problems became the only opportunity to earn money.

Methods. We recruited 201 adult participants on Amazon Mechanical Turk. Participants were paid $0.75 for participation and could earn a performance-dependent bonus of up to $2. After performing the task participants completed an attention check that required them to estimate the highest possible payoffs of the different types of games they played in the experiment. Participants were excluded if they reported a positive number for the gamble that had only negative outcomes, if their estimate for the high-stakes gamble (± 100) was less than twice their estimate for the low-stakes gamble (± 10), or if any of their estimates was larger than 500. Based on these criteria, we had to exclude 57 participants (28.36%). In the experiment, participants visited a virtual casino that offered three different kinds of games: In *Blue Mountain* Games the stakes were high (± 100). In *Purple Sun* Games the stakes were low (± 10), and in *Orange Diamond* games all outcomes were negative $([-100; -90])$. Each type of game was associated with a logo. The instructions informed participants that there were three kinds of games and what their payoffs were. In contrast to Experiment 2, the range of possible outcomes was not stated explicitly on every trial; instead they had to be inferred from the game's logo. Figure 8 shows a screenshot from Experiment 3.



Lucky Day Casino

Round #1 Choice #1 0 Points

01 37
MINUTES SECONDS

Please click on the gamble you would like to play.

Balls	Gamble 1	Gamble 2	Gamble 3	Gamble 4	Gamble 5
11 x A	-88	?	?	?	?
5 x B	?	?	?	?	?
42 x C	?	?	?	?	?
42 x D	?	?	?	?	?

Figure 8: Screenshot from Experiment 3.

The experiment was structured into five blocks lasting 2.5 minutes each. The first and the second block had a high reward rate. They comprised 50% high-stakes problems and 50% low-stakes problems. Blocks 3-5 were the pretest, the training, and the posttest block respectively, and they all had the same structure. In each of these blocks, the first four trials were low-stakes problems (± 10 points) and the remaining trials comprised 75% trials with only negative outcomes ($[-100, -90]$) and 25% low-stakes problems. Hence, starting with the pretest block, low-stakes decisions became the only opportunity to win points and the opportunity cost for engaging in them became negligible. In all decision problems presented in Experiment 3, identifying the optimal choice required integrating multiple attributes. The number of gambles was always five, and the number of outcomes was always four. The payoffs were sampled uniformly from the range associated with the problem and their probabilities were determined as in Experiment 2. In contrast to Experiment 2, participants could *not* skip trials but always had to choose a gamble to advance.

We ran 200 simulations of this experiment using the same strategies and parameters as for Experiment 2 except that the agent did not have the option to skip trials. Rational metareasoning predicted that starting from the pretest (block 3), participants will learn to reap increasingly higher reward rates by engaging more often in the now worthwhile low-stakes problems and acquire increasingly more information to make those choices (see Figure 9).

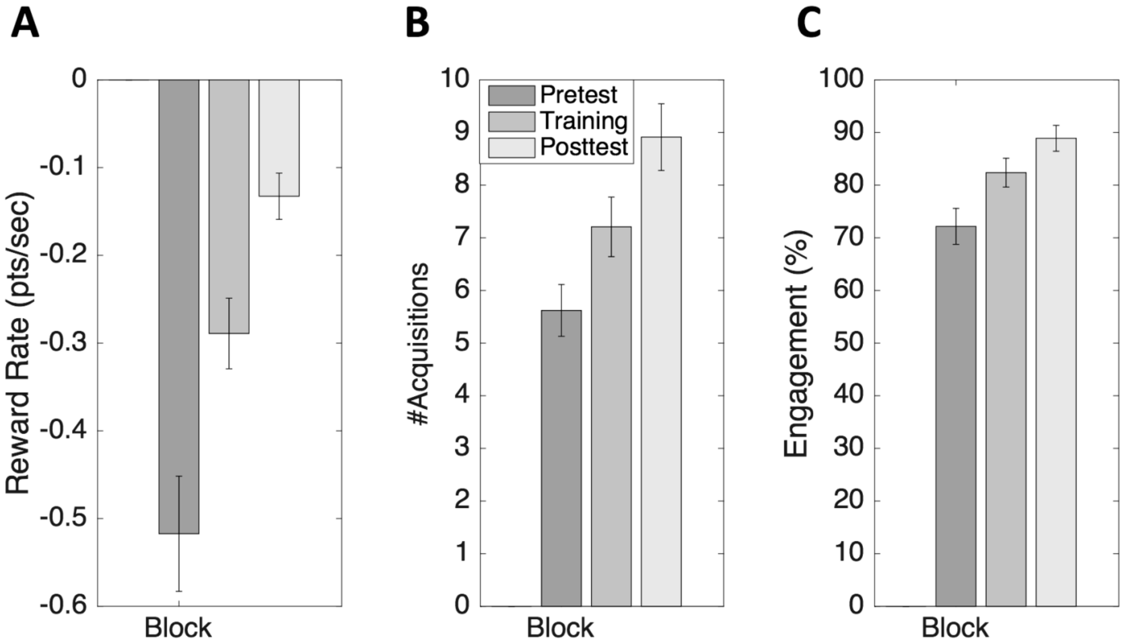


Figure 9: Rational metareasoning predictions of strategy selection learning in Experiment 3. A: Rational metareasoning predicted a significant increase in the reward rate from the pretest (block 3) to the posttest (block 5). B: Rational metareasoning predicted a significant increase in the number of information acquisitions on the worthwhile low-stakes problems. C: Rational metareasoning predicted a significant increase in people's engagement with the worthwhile low-stakes problems.

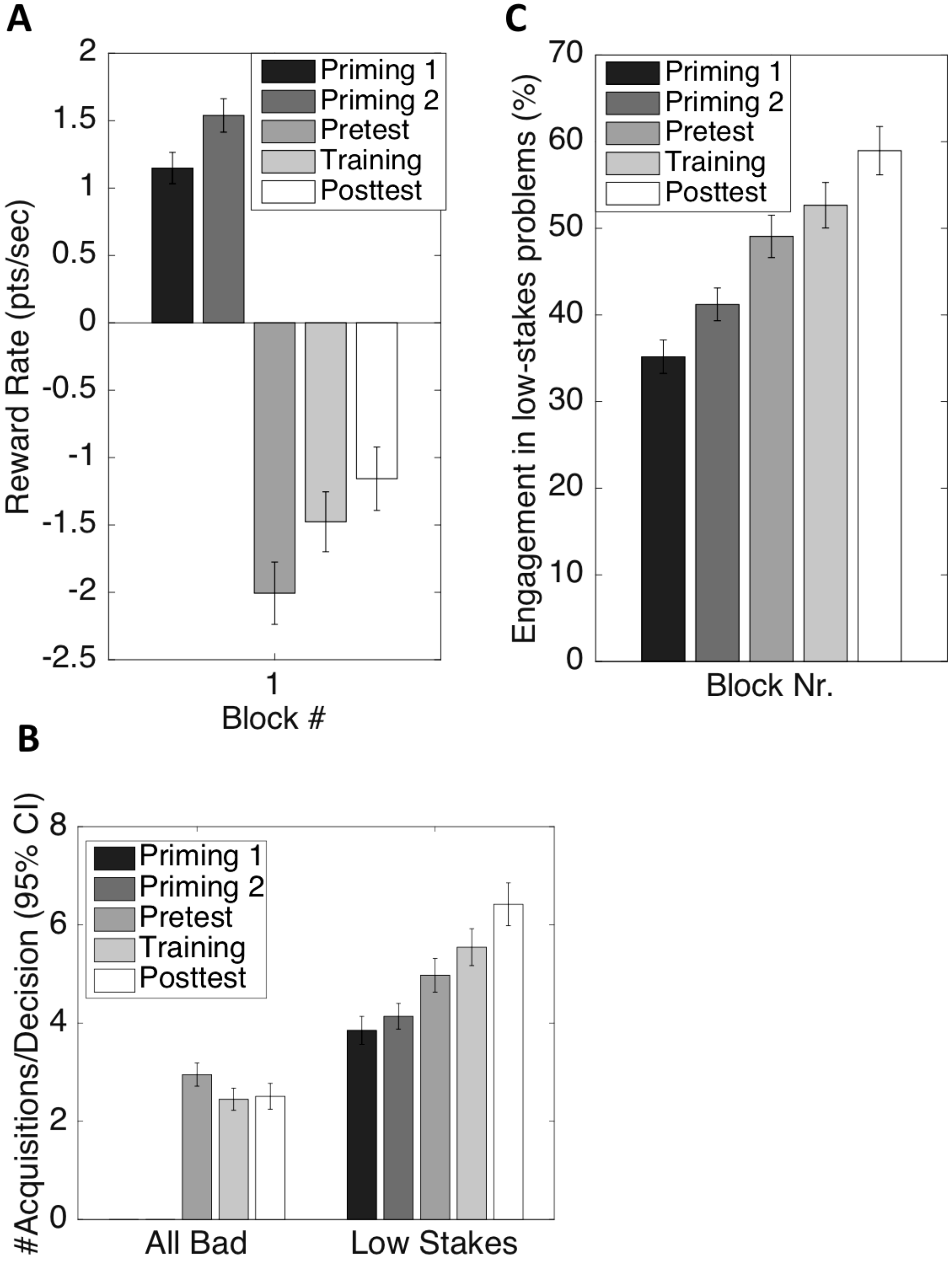


Figure 10: Empirical Results of Experiment 3. A: Reward rate by block. Error bars denote plus and minus one SEM. B: Average number of information acquisitions. C: Engagement in low-stakes decisions.

Results and Discussion. First, we quantified learning by the change in our participants' average reward rate from the pretest to the posttest. The increase in people's average reward rate from -2.00 ± 0.24 in the pretest to -1.16 ± 0.23 in the posttest was statistically significant according to a one-sided t-test ($t(143) = 2.87, p = 0.002$; Cohen's $d = 0.26$). The reward rate depends on two factors: the reward per decision and the number of decisions per minute. To determine which of the two factors was responsible for the increase, we analyzed the learning induced changes in each factor separately. First, we analyzed how the reward per decision changed from the pretest to the posttest. For low-stakes problems the reward increased significantly from about 1.93 points per decision to about 2.42 points per decision ($t(143) = 1.92, p = 0.03$; Cohen's $d = 0.16$). By contrast, for problems on which all outcomes were negative the average reward did not change significantly (-14.43 vs. $-14.22, t(96) = 1.08, p = 0.14$; Cohen's $d = 0.11$). Next, we analyzed potential changes in the second factor: the number of decisions per unit time. We found that participants slowed down significantly from 12.86 ± 1.42 decisions per minute in the pretest to 8.32 ± 1.45 decisions per minute in the posttest ($t(143) = 2.99, p = 0.003$; Cohen's $d = 0.25$). Hence, participants learned to reap a higher reward rate by deliberating more to make better decisions.

To test the hypothesis that deliberation increased with learning more rigorously, we analyzed the number of information acquisitions as a proxy for the number of computations performed by our participants. Concretely, we tested our model's prediction that people should learn to invest more computation into low-stakes decisions. As shown in Figure 10A, participants learned to allocate their time adaptively. Starting from the pretest (block 3)—where low-stakes problems became worthwhile solving—there was a significant increase in the number of information acquisitions on the low-stakes problems from 4.97 ± 0.34 to 6.42 ± 0.43 ($t(2798) = 5.19, p < 0.001$; Cohen's $d = 0.10$). This increase was specific to the low-stakes problems: It did not occur for problems with only negative outcomes. To the contrary, on problems with only negative outcomes the number of information acquisitions decreased from 2.95 ± 0.24 to 2.51 ± 0.26 ($t(5112) = -2.38,\ p = 0.02$; Cohen's $d = 0.03$). This suggests that people learned to allocate their computation more adaptively from the pretest to the posttest. The number of information acquisitions was particularly high on the first four trials of the three last blocks: the number of information acquisitions increased from 8.19 ± 0.51 in the pretest to 9.27 ± 0.50 in the posttest ($t(143) = 2.14,\ p = 0.02$; Cohen's $d = 0.18$).

The observed increase in the number of information acquisitions on low-stakes problems might be caused by an increase in the frequency with which people engaged with them, an increase in the number of computations they invested into solving those they engaged with, or both. We found that the increase in the number of information acquisitions per problem was mostly driven by an increase in the frequency with which people engaged in effortful decision making on low-stakes problems (see Figure 10B): the frequency of engagement in low-stakes problems increased from only $49.1 \pm 0.2\%$ in the pretest to $58.95 \pm 2.8\%$ in the posttest ($\chi^2(2) = 26.9, p < 0.001$; Cohen's $w = 5.19$). This increase was accompanied by a decrease in the frequency with which participants chose randomly which was the only way to avoid engaging with the problem. Importantly, we also found that the number of inspected outcomes increased even on the low-stakes problems that participants engaged with (10.14 in the pretest versus 10.89 acquisitions in posttest, $t(1490) = 2.07,\ p = 0.04$; Cohen's $d = 0.05$). On the problems with only negative outcomes, by contrast, there was a significant decrease in the number of information acquisitions ($t(1291) = -8.52, p < 0.001$; Cohen's $d = -0.24$). In conclusion, the increase in the number of information acquisitions on low-stakes problems was driven by both factors: our participants learned to engage in low-stakes decisions more frequently and to deliberate more when engaged. Both changes are consistent with learning to become more resource-rational. Finally, we also found that people gradually learn to prioritize the most probable outcomes. The average rank of the outcome that participants inspected first significantly decreased from 2.36 ± 0.07 in the first block to 2.18±0.11 in the last block ($t(133) = 3.96,\ p < 0.001$; Cohen's $d = 0.34$). This learning process occurred even though identifying the optimal decision always required inspecting multiple outcomes.

As for Experiment 2, the predictions of our rational model were qualitatively correct, but the observed learning effects were slightly smaller than expected. The model achieved a slightly higher reward rate than people (cf. Figure 9A vs. Figure 10A), acquired about 0.5—3 additional pieces of information (cf. Figure 9B vs. Figure 10B), and engaged in 20%—30% more problems than people (cf. Figure 9C vs. Figure 10C). In summary, we found that people learn to deliberate more and gather more information when the reward structure of their environment calls for it. This result complements the finding from Experiment 2 where people learned to invest less computation because the return on investing deliberation was less than their opportunity cost. In conclusion, our results suggest that strategy selection learning makes people more resource-rational by tuning strategy choices towards the optimal speed-accuracy tradeoff.

Model Comparisons. For the purpose of model comparison, we ran 200 simulations of Experiment 3 according to each of the 14 alternative models described above. For each model, we performed six one-sample t-tests to determine whether it correctly predicted the increases in reward rate, information acquisitions, and the frequency of engagement that occurred from block 3 to block 4 and from block 4 to block 5, as well as one t-test to evaluate whether the model captured that people gathered more information on high-stakes problems than on other kinds of problems. We found that while our rational metareasoning model captured all of these effects, none of the SCADS, RELACS, or SSL models was able to capture these four effects simultaneously. Among the lesioned metareasoning models, only the one approximating the VOC by model-free reinforcement learning from the difference between reward and time cost captured all four phenomena. Critically, none of the other lesioned metareasoning models were able to do so. This suggests that choosing strategies based on the VOC, exploration, and feature-based learning are necessary to capture the adaptive strategy selection learning our participants demonstrated in Experiment 3. Hence, only the full rational metareasoning model can capture the findings from Experiments 2 and 3 simultaneously. For a more detailed summary of these simulation results, please see Appendix B.

According to our rational theory of strategy selection, the reason why some people are cognitive misers in certain tasks (Toplak, West, & Stanovich, 2013) is that their metacognitive model predicts that the reward for normative performance is just not worth the effort it would require. The results of Experiment 2 suggest that cognitive misers will often learn to deliberate more when the returns of deliberation justify its cost.

Ecological rationality increases with learning

The third prediction of our model is that people adapt their strategy choices to the structure of their environment. To evaluate this prediction, we examined a concrete example where people can use two different strategies to choose between two options with multiple attributes[5]: the comprehensive Weighted-Additive-Strategy (WADD) versus the fast-and-frugal heuristic Take-The-Best (TTB). There are different variants of the WADD strategy. Since we will be modeling a multi-attribute binary choice task, we use the version of WADD that sums up the weighted differences between the first option's rating and the second option's rating across all attributes (Tversky, 1969). For each attribute this strategy compares the two ratings (1 operation). If the attribute values disagree, then it reads and adds or subtracts the attribute's validity (2 operations). Finally, it chooses the first attribute if the sum is positive or the second attribute if the sum is negative (1 operation). TTB is the equivalent of the lexicographic heuristic for multi-attribute decisions: it chooses the option that is best on the most predictive attribute that distinguishes between the options and ignores all other attributes. Our implementation of Take-The-Best first searches for the most predictive attribute by sequentially reading the validities of unused attributes (1 operation per attribute), comparing them to the highest validity found so far (1 operation per attribute), and memorizes the new validity if it exceeds the previous maximum (1 operation). Once the most predictive attribute has been identified, TTB compares the options' ratings on that attribute (1 operation), and then either makes a choice (1 operation), or continues with the next most predictive attribute.

TTB works best when the attributes' predictive validities fall off so quickly that the recommendation of the most predictive attribute cannot be overturned by rationally incorporating additional attributes; environments with this property are called *non-compensatory*. By contrast, TTB can fail miserably when no single attribute reliably identifies the best choice by itself; and environments with this property are called *compensatory*. Thus, to adapt rationally to the structure of their environment, that is to be *ecologically rational*, people should select TTB in non-compensatory environments and avoid it in compensatory environments.

Bröder (2003) found that people use TTB more frequently when their decision environment is non-compensatory. Rieskamp and Otto (2006) found that this adaptation might result from reinforcement learning. In their experiment participants made 168 multi-attribute decisions with feedback. In the first condition, all decision problems were compensatory, whereas in the second condition all decision problems were non-compensatory. To measure people's strategy use over time, Rieskamp and Otto (2006) analyzed their participants' choices on trials where TTB and WADD made opposite decisions. Participants in the non-compensatory environment learned to choose in accordance with TTB increasingly *more* often, whereas participants in the compensatory environment learned to do so increasingly *less* often.

These findings raise the question of how people learn when to use TTB. One hypothesis is that people learn how well TTB works *on average*, as postulated by the SSL model (Rieskamp & Otto, 2006). Our alternative hypothesis is that people learn to predict how fast and how accurate TTB and alternative strategies will be on *individual problems* based on problem features, as postulated by rational metareasoning. To test these two hypotheses against each other, we simulated Experiment 1 from Rieskamp and Otto (2006) according to rational metareasoning and SSL and compared how well the models' predictions explained the data. The experiment was divided into seven blocks. Each block comprised 24 trials, and each trial presented a choice between two investment options

[5] A preliminary version of these simulations appeared in Lieder and Griffiths (2015).

with five binary attributes. The attributes' predictive validities were constant and explicitly stated. Both models assumed that participants in this experiment always choose between Take-The-Best ($s_1 = \mathrm{TTB}$) and the weighted-additive strategy ($s_2 = \mathrm{WADD}$). Our rational metareasoning model of this paradigm assumed that strategy selection in binary multi-attribute decisions relies on three features $\mathbf{f} = (f_1, f_2, f_3)$: the predictive validity of the most reliable attribute that discriminates between the two options (f_1), the gap between the validity of the most reliable attribute favoring the first option and the most reliable attribute favoring the second option (f_2), and the absolute difference between the number of attributes favoring the first option and the second option respectively (f_3). Our model assumes that people first inspect the validities of all cues and extract the three features $f_1, \cdots, f_3$ from them, then select a strategy based on these features, and finally execute that strategy to reach a decision.[6]

The probability that a strategy s makes the correct decision ($R = 1$) was modeled by

$$P(R = 1|s, \boldsymbol{f}) = \frac{1}{1 + \exp\left(-\left(b_s + \sum_i w_{s,i} \cdot \mathrm{f_i}\ \right)\right)}.$$

We modeled people's knowledge about the feature weights w_s by the prior distribution

$$P(w_s) = \mathcal{N}\left(\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma^{-1} = \tau \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right),$$

$$P(b_s) = \mathcal{N}\left(\mu = b_s^{(0)}, \sigma^{-2} = \tau\right),$$

where the expected value of the offset b_s (i.e., $b_s^{(s)}$) and the strength τ of the prior belief are free parameters.

To simulate the first experiment from Rieskamp and Otto (2006), we created a compensatory environment and a non-compensatory environment. In the compensatory environment, WADD always makes the Bayes-optimal decision and TTB disagrees half of the time. Conversely, in the non-compensatory environment TTB always makes the Bayes-optimal decision and WADD disagrees half of the time. To determine the optimal choices, we computed the probability that option A is superior to option B given their ratings by Bayesian inference under the assumption that positive and negative ratings are equally common. First, we randomly generated a set of candidate decision problems. For each of these problems, we computed the posterior probability that the first option is superior to the second option given their attributes' values and their validities. We then used these posterior probabilities to select which candidate decision problems to present in the compensatory environment and which to present in the non-compensatory environment. To match the reward probabilities of Experiment 1 by Rieskamp and Otto (2006), the feedback was determined solely based on the environment and the chosen strategy: the probability of positive feedback was 92% whenever the strategy matched the structure of the environment (e.g., WADD in the compensatory environment) and only

[6] This entails that all cue validities are inspected on all trials even when a fast-and-frugal heuristic like TTB is chosen. This makes the number of information acquisitions on trials where TTB is used more similar to the number of information acquisitions on trials where WADD is used. This diminishes the relative number of information acquisitions saved by TTB. However, this does *not* affect the number of inspected cue values.

58% when it did not (e.g., TTB in the compensatory environment). Positive feedback meant winning \$0.15 whereas negative feedback meant losing \$0.15.

To simulate the experiment, we let our rational metareasoning models learn the agent's opportunity cost from experience; the prior mean of the opportunity cost was initialized with \$7/h and the prior precision corresponded to one minute's worth of experience. For simplicity, we assumed that people perform one step of TTB or WADD per second. To estimate which strategy people considered more effective a priori, we set the prior expectation of the problem-independent performance of TTB ($b_{\text{TTB}}^{(0)}$) to zero and fit the model's prior expectation of the problem independent performance of WADD ($b_{\text{WADD}}^{(0)}$) and the strength of the agent's prior beliefs about the strategies' performance and execution time (τ) to the data. Specifically, we determined these parameters by maximum-likelihood estimation from the frequencies with which Rieskamp and Otto's participants used TTB in each block using grid search. The likelihood function was estimated by running at least 10 simulations of the experiment for each point on the grid of potential parameter values. Rieskamp and Otto (2006) estimated that participants made accidental errors in about 5% of the trials. To capture these errors and avoid numerical problems, we modelled people's apparent strategy choice frequencies by

$$\hat{\theta}_{\text{strategy}}^{(b)} = \frac{0.9 \cdot n_{\text{strategy}}^{(b)} + 0.1 \cdot 0.5 \cdot n_{\text{total}}^{(b)}}{n_{\text{total}}^{(b)}} \qquad (5),$$

where strategy is a placeholder for either TTB or WADD and $n_{\text{total}}^{(b)} = n_{\text{TTB}}^{(b)} + n_{\text{WADD}}^{(b)}$ is the total number of trials in block b.[7]

The resulting parameter estimates captured that people initially preferred WADD to TTB ($\hat{b}_{\text{WADD}}^{(0)} = +0.32$) and required many decisions' worth of experience to revise their beliefs ($\hat{\tau} = 88.59$). Our simulation showed that rational metareasoning can explain people's ability to adapt their strategy choices to the structure of their environment (see Figure 11): When the decision environment was non-compensatory, then our model learned to use TTB and avoid WADD. But when the decision environment was non-compensatory, then our model learned to use WADD and avoid TTB. In addition, rational metareasoning captured that people adapt their strategy choices gradually.

We also estimated the parameters of the SSL model and the three SCADS models introduced above. The SCADS models were equipped with two categories for compensatory versus non-compensatory problems respectively. The free parameters of the SCADS models determined the initial associations between each category and the two strategies. The first parameter was the sum of the two strategies' association strengths, and the second parameter was the relative strength of the association with the WADD strategy. The global association strengths were the sums of the category-specific associations. For the SSL model, we estimated the relative reward expectancy of the

[7] This assumption is not a model of the underlying psychological processes. Instead, it serves as a placeholder for all unknown and known influences on strategy selection that the model does not capture. The frequency of trials in which the strategy is chosen at random was selected so as to generate 5% of trials in which the chosen strategy disagrees with the one prescribed by the model. We assumed random choice because it is the weakest assumption we could make.

WADD strategy (β_{WADD}) and the strength of the initial reward expectancy (w) by the simulation-based maximum-likelihood method described above (Equation 5).

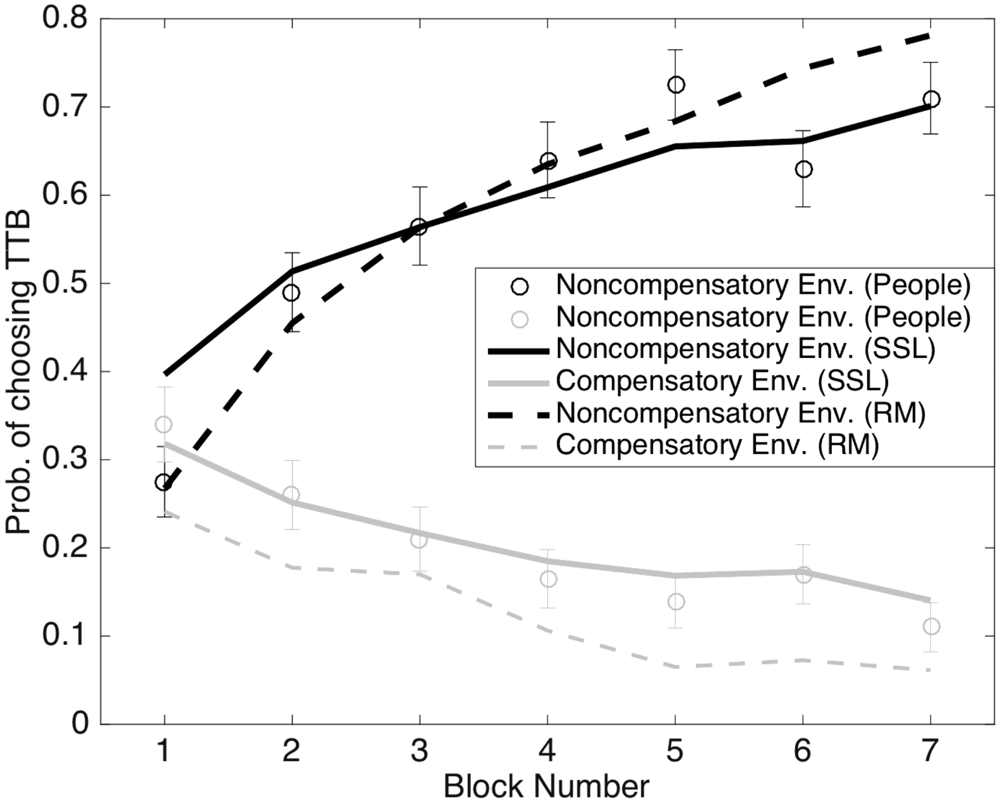


Figure 11: Fit of rational metareasoning model and SSL to the empirical data by Rieskamp and Otto (2006).

The maximum-likelihood estimates of the SSL model's parameters were $\hat{\beta}_{\text{WADD}} = 0.35$ and $\hat{w} = 30$. The mean squared error of the fit achieved by the SSL model was about half the MSE of the rational metareasoning model (0.0018 vs. 0.0043)' see Figure 11. Consequently, the Bayesian information criterion provided strong evidence for the SSL model over the full rational metareasoning model (BIC_{SSL}=60.70 vs. BIC_{RM}=68.94; Kass & Raftery, 1995) and all lesioned metareasoning models ($\text{BIC} \geq 70.52$). The BIC of the full rational metareasoning model was slightly higher than the BIC for the lesioned metareasoning model without features ($\text{BIC} = 70.52$), and the data provided strong or very strong evidence for the full metareasoning model over all other lesioned metareasoning models ($\text{BIC} \geq 75.94$). The fit of the SCADS models was comparable to the fit of the SSL model and significantly better than the fits of the metareasoning models ($\text{BIC}_{\text{SCADS1}}$ = 61.09, $\text{BIC}_{\text{SCADS2}}$ = 62.38, and $\text{BIC}_{\text{SCADS3}}$ = 62.01). Finally, we repeated our model comparison for both environments separately. Consistent with the original model comparison results, we found that SSL provided a better explanation for the data from the compensatory environment ($\text{BIC}_{\text{SSL}} = 32.64$ vs. $\text{BIC}_{\text{RM}} = 34.29$) and the data from noncompensatory environment ($\text{BIC}_{\text{SSL}} = 36.22$ vs. $\text{BIC}_{\text{RM}} = 37.81$) than the rational metareasoning models. The performance of the SCADS models was close to the performance of the SSL models ($\text{BIC}_{\text{SCADS}} = 32.72$ for the compensatory environment and $\text{BIC}_{\text{SCADS}} = 36.35$ for the noncompensatory environment).

The quantitative differences between the model fits should be taken with a grain of salt because they depend on the auxiliary assumption that people use the exact TTB and WADD strategies available to the models and no other strategies. This assumption is questionable for at least two reasons: First, TTB and WADD are merely placeholders for the class of non-compensatory strategies and the class of compensatory strategies respectively (Rieskamp & Otto, 2006). Second, previous work suggests that the human mind is equipped with a much larger repertoire of decision strategies (Payne, et al., 1988). If the rational metareasoning model was also equipped with a larger repertoire of strategies, then it would learn more gradually and probably achieve a better fit to the human data. Due to these caveats, we focus on the models' qualitative predictions because they are less sensitive to different auxiliary assumptions.

The feature-based learning mechanism of the SCADS model and the context-free learning mechanism of the SSL model captured the human data equally well (BIC_{SSL} -$\text{BIC}_{\text{SCADS1}} = 0.04 \ll 2$), and the feature-based learning mechanism of the rational metareasoning model also captured the qualitative changes in people's strategy choices. Since the data by Rieskamp and Otto (2006) can be explained by either feature-based or context-free strategy selection learning, we designed a new experiment to determine which mechanism is responsible for people's adaptive strategy choices.

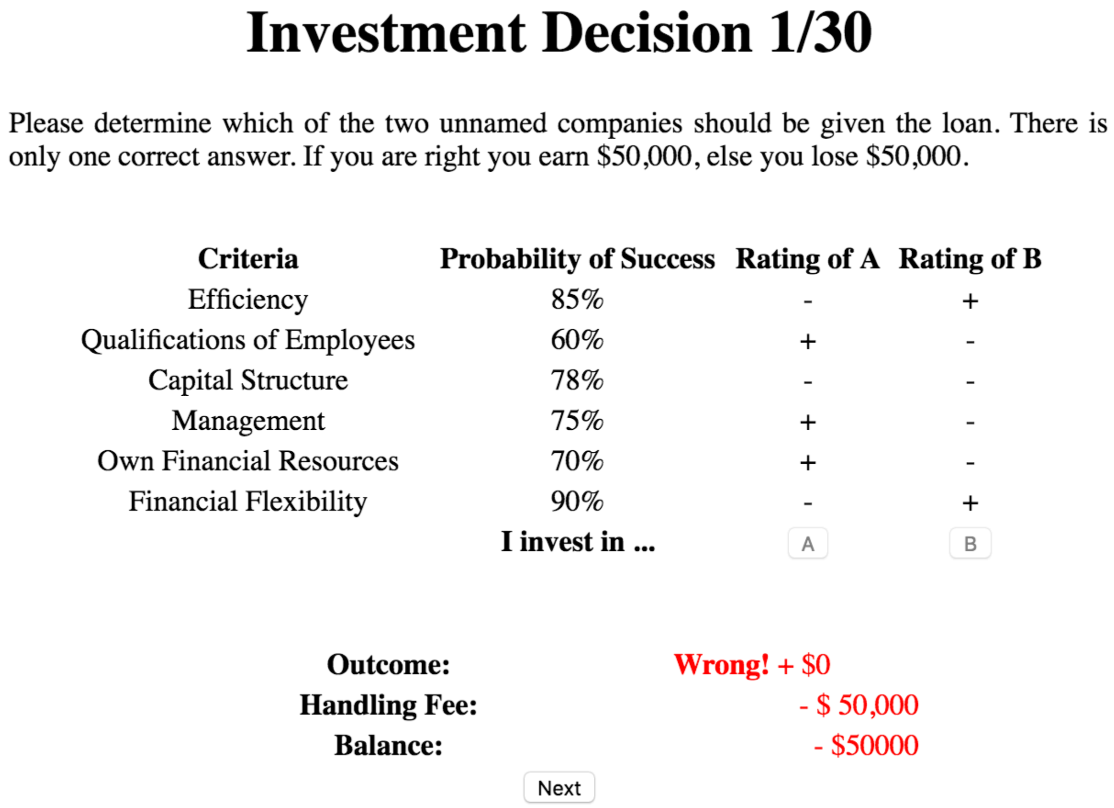
Experiment 4: Adaptive flexibility increases with learning

The fourth prediction of our model is that people learn to flexibly switch their strategies on a trial-by-trial basis to exploit the structure of individual problems. An alternative hypothesis embodied by SSL and RELACS is that strategy selection learning serves to identify the one strategy that works best on average across all problems in a given environment. To design an experiment that can discriminate these hypotheses, we evaluated the performance of context-free versus feature-based strategy selection learning in 11 environments with $0\%, 10\%, 20\%, \cdots, 100\%$ compensatory problems and 100%, $90\%,\ 80\%,\ \cdots, 0\%$ non-compensatory problems respectively. Critically, all compensatory problems were designed such that TTB fails to choose the better option and WADD succeeds, and all non-compensatory problems were designed such that TTB succeeds and WADD fails. For each of the 11 decision environments, we compared the average performance predicted by rational metareasoning with the parameters $b_{\text{TTB}}^{(0)} = b_{\text{WADD}}^{(0)} = 0$ and $\tau = 1$, against the predictions of the five lesioned metareasoning models with the same parameters, SSL with parameters $\beta_1 = \beta_2 = 0.5$ and $w = 1$, RELACS with parameters $\alpha = 0.1$ and $\lambda = 1$, and the three SCADS models with an association strength of 0.5 between each strategy and two categories corresponding to compensatory and non-compensatory problems respectively.[8]

Our simulations revealed that feature-based and context-free strategy selection learning predict qualitatively different effects of the relative frequency of compensatory

[8] These parameters were chosen to give each model a weak, initial bias towards using both strategies equally often. The exact value of this bias is not critical because it is quickly overwritten by experience.

versus non-compensatory decision problems; see Figure 13A. Concretely, the performance of model-free strategy selection learning drops rapidly as the decision environment becomes more heterogeneous: As the ratio of compensatory to non-compensatory problems approaches 50/50 the performance of context-free strategy selection learning (SSL, RELACS, and the lesioned metareasoning model without features) and SCADS[9] drops to the level of chance. By contrast, the performance of feature-based strategy selection learning (rational metareasoning) is much less susceptible to this heterogeneity and stays above 60%. The reason is that rational metareasoning learns to use TTB for non-compensatory problems and WADD for compensatory problems, whereas SSL and RELACS learn to always use the same strategy. We can therefore determine whether people rely on context-free or feature-based strategy selection with the following experiments that puts participants in a heterogeneous environment.



Investment Decision 1/30

Please determine which of the two unnamed companies should be given the loan. There is only one correct answer. If you are right you earn $50,000, else you lose $50,000.

Criteria	Probability of Success	Rating of A	Rating of B
Efficiency	85%	-	+
Qualifications of Employees	60%	+	-
Capital Structure	78%	-	-
Management	75%	+	-
Own Financial Resources	70%	+	-
Financial Flexibility	90%	-	+
	I invest in ...	A	B

Outcome:	**Wrong!** + $0
Handling Fee:	- $ 50,000
Balance:	- $50000

Figure 12: Interface of Experiment 4: Strategy selection in multi-attribute decision-making.

Methods. We recruited 100 participants on Amazon Mechanical Turk. The experiment lasted about 25-30 min, and participants were paid $1.25 plus a performance-

[9] The problem preventing the SCADS model from choosing the best strategy for each category is that the category-specific association strengths are multiplied by a category-unspecific association strength.

dependent bonus of up to \$1.25. The experiment instantiated the decision environment with 50% compensatory problems and 50% non-compensatory problems from the simulations above. Participants played a banker deciding between giving a loan to company A versus company B based on their ratings on multiple attributes with explicitly stated predictive validities (see Figure 12). There were 12 attributes in total. Half of these attributes were reliable (predictive validity $\geq 85\%$) whereas the other half was unreliable (predictive validity $\leq 63\%$). Concretely, the attributes *Financial Flexibility, Efficiency, Capital Structure, Management, Own Financial Resources, and Qualifications of Employees* had predictive validities of 95%,93%,90%,87%,85%, and 83% respectively, whereas the attributes *Investment Policy, Business History, Real Estate, Industry, Reputation*, and *Location* had predictive validities of 63%, 60%, 57%, 55%, 53%, and 51% respectively. Each trial presented either 3, 4, 5, or 6 attributes with equal probability. On non-compensatory trials, exactly one of the attributes was reliable and all other attributes were unreliable. By contrast, on compensatory trials all attributes were reliable or all attributes were unreliable. Reliable and unreliable attributes were selected randomly and their order was randomized. The two options always had opposite ratings on the most predictive attribute, and 75% of the ratings on other attributes were opposite to the rating on the most predictive attribute while 25% agreed with it. After choosing *Company A* or *Company B,* participants received stochastic binary feedback: \$+50,000 versus \$-50,000. On compensatory trials, the probability of positive feedback was 95% when the participant's choice agreed with the choice of WADD and 5% when it disagreed with WADD. On non-compensatory trials the probability of positive feedback was 95% when their choice agreed with TTB and 5% otherwise.

Each participant made 100 binary choices, earning a bonus of 1.25 cents for each correct decision and losing 1.25 cents for each incorrect decision. Critically, the ratio of compensatory to non-compensatory problems was 50/50: The problems were chosen such that TTB and WADD make opposite decisions on every trial. In half of the trials, the decision of TTB was correct and in the other half WADD was correct. Therefore, always using TTB, always using WADD, choosing one of these two strategies at random, or context-free strategy selection would perform at chance level; see Figure 13A.

Results and Discussion. To determine the quality of people's strategy choices, we compared their decisions on each trial to those of the strategy appropriate for the problem presented on that trial. For compensatory trials, we evaluated people's choices against those of WADD and for non-compensatory trials we evaluated them against TTB. People's decisions agreed with those of the appropriate strategy on 76.2% of the trials (see Figure 13A). To quantify our uncertainty about this estimate, we computed its credible interval assuming a uniform prior (Edwards et al., 1963). We found that the 99% highest-posterior density interval ranged from 75.1% to 77.3%. We can thus be 99% confident that people's average performance in the mixed decision environment was at least 75% and conclude that they performed significantly better than chance ($p < 0.001$, Cohen's $w = 52.36$). As shown in Figure 13B, people's performance increased significantly from 70.4% in the first ten trials to 80.4% in the last ten trials ($\chi^2(1) = 26.96, p < 0.001$, Cohen's $w = 5.19$). To gain a better understanding of this effect, we performed a logistic regression of the agreement between people's choices and those of the appropriate strategy; the regressors were the trial number, a constant, and the decision's compensatoriness. We found that people's performance increased significantly over trials ($t(9996) = 9.46, p < 0.001$). Consistent with the finding that people initially prefer compensatory strategies (Rieskamp & Otto, 2006), people performed better on compensatory trials than on non-compensatory trials overall ($t(9996) = 9.46, p < 0.001$) and this effect dissipated over time ($t(9996) = -7.20, p < 0.001$). Analyzing compensatory and non-compensatory trials separately with logistic regression revealed that our participants' performance on non-compensatory trials improved significantly over time ($t(4998) = 9.46, p < 0.001$) while their performance on compensatory trials remained constant ($t(4998) = -0.92, p = 0.36$). Interestingly, people performed significantly above chance already on the first trial (73% correct; $p < 0.001$). This suggests that people either entered the experiment with applicable expertise in when to use compensatory versus non-compensatory decision strategies, as suggested by the results of Payne et al. (1988) or possess general purpose strategies that work well on both kinds of problems. Both factors might also explain why people performed systematically better than all of our models (Figure 13A).

This level of performance could not have been achieved by context-free strategy selection, which performed at chance, but it is qualitatively consistent with feature-based strategy selection which performed significantly better than chance; see Figure 13A. We also simulated the experiment with three SCADS models that were equipped with two categories corresponding to compensatory versus non-compensatory problems and differed in their reward function ($r = \text{correctness}$, vs. $r = \text{correctness} - \text{cost}$, vs. $r = \text{correctness/time}$). We found that the performance of the SCADS models was very similar to the performance of the SSL model. Most importantly, its performance dropped to the chance level as the environment became increasingly more heterogeneous. This happened because the global association strength interfered with category-specific strategy choices. Additional simulations with the five lesioned metareasoning models revealed that feature-based learning was indispensable to capture human performance. For more information, see Supplementary Figure 9 in the Supplementary Online Material.

These results should be taken with a grain of salt because the model comparisons presuppose that TTB and WADD are the only decision strategies that people are equipped with even though people's repertoire most likely includes many additional strategies. It is conceivable that participants succeeded in Experiment 4 by relying on a single strategy that succeeds on both compensatory and non-compensatory problems. Because of this possibility, Experiment 4 does not provide definite evidence for feature-based strategy selection. However, Experiment 1, Experiment 3, and the simulations of mental arithmetic presented in the following sections also support feature-based strategy selection. Taken together these experiments and simulations provide very strong evidence for feature-based strategy selection learning.

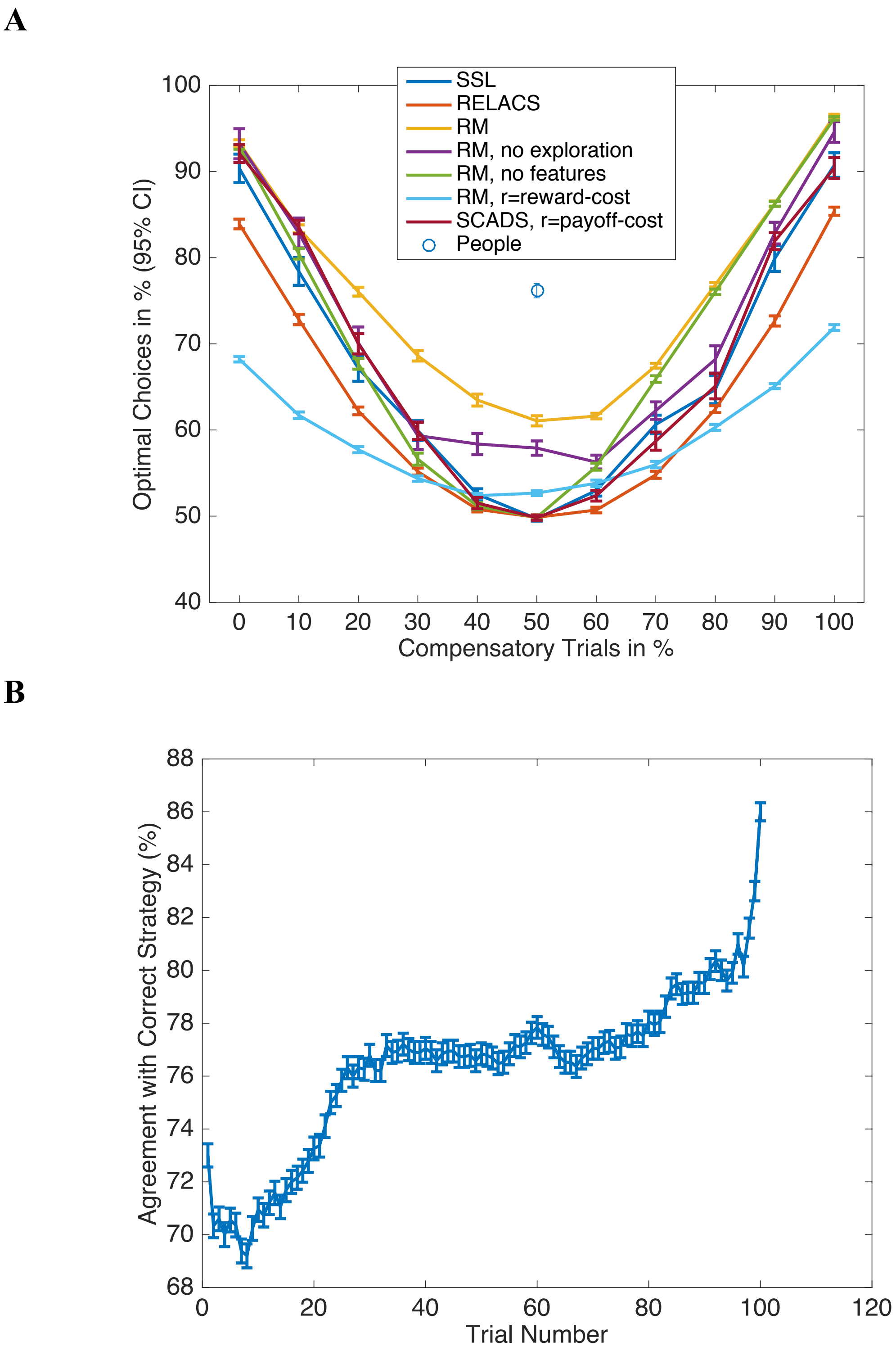


Figure 13: Model predictions and findings of Experiment 4. A: People and rational metareasoning perform significantly above chance in heterogeneous environments but context-free strategy selection mechanisms do not. B: People's performance increased with experience. The trial-by-trial frequencies were smoothed by a moving average over 20 trials. The error bars enclose 95% confidence intervals.

Conclusion

The experiments presented in this section confirmed the predictions of our resource-rational theory of strategy selection learning: The first experiment showed that people learn to think less when they think too much. The second experiment showed that people learn to think more when they think too little. Thirdly, we showed that people learn to adapt not only how much they think but also *how* they think to the structure of the environment. Finally, Experiment 4 demonstrated that adaptive flexibility also increases with learning, and this enables people's strategy choices to exploit the structure of individual problems. Most importantly, in all four cases, the underlying learning mechanisms made people's strategy choices increasingly more resource-rational. Hence, the empirical evidence presented in this section supports our hypothesis that the human brain is equipped with learning mechanisms that make it more resource-rational over time. Even though people may not be resource-rational when they first enter a new environment, the way in which they process information appears to converge to the rational use of their finite time and bounded computational resources. Given the support for this view in the domain of decision-making, the following two sections investigate whether this conclusion also holds for other domains.

Strategy selection and cognitive development

So far, we have found that adults' strategy choices in sorting, decision-making, and planning become increasingly more rational through learning within minutes. Since learning is an important driving force of cognitive change our theory predicts similar phenomena should also occur on the much longer time-scales of cognitive development.

A substantial literature on the development of children's arithmetic competencies suggests that cognitive development does not proceed in a sequence of discrete stages characterized by a progression of beliefs, representations, and cognitive strategies as proposed by Piaget (Piaget & Cook, 1952) but rather as a gradual shift in the frequency with which children use each of multiple coexisting cognitive strategies (Siegler & Shipley, 1995). According to Siegler's *overlapping waves* theory of cognitive development (see Figure 14A; Siegler, 1996) children of every age use a variety of strategies, and over time strategies that are both effective and efficient come to be used more frequently.

To give just one example of such *strategic development* (Siegler, 1999) we consider the development of children's strategies for mental addition shown in Figure 14B. Svenson and Sjöberg (1983) found that the Retrieval strategy becomes increasingly more prominent as children get older, while the frequency of not providing an answer drops rapidly. The frequency of the *Sum* strategy rises initially making it the most common strategy at the beginning of second grade, but afterwards its frequency drops again. The frequency of the *Min strategy* rises initially, and then it stays roughly constant until the Min strategy is overtaken by the *Retrieval strategy*.

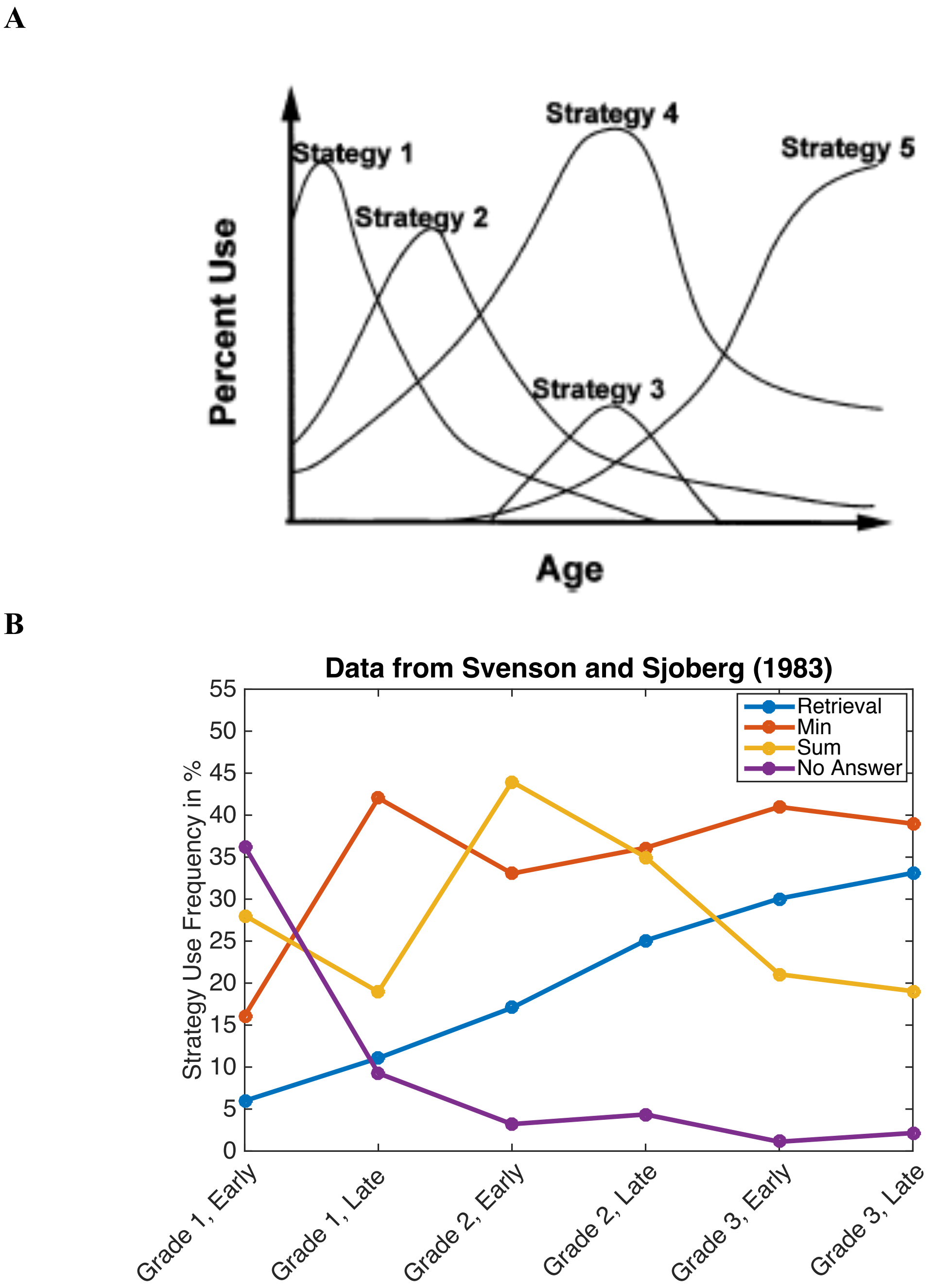


Figure 14: Overlapping waves theory of cognitive development . (A) Illustration of the theory from Siegler (1999). (B) Empirical support for overlapping waves in the development of children's strategy use in mental addition according to Svenson and Sjobert (1983).

Children's strategic development raises the question of how they learn to use effective strategies more and less effective strategies less. Learning to use effective strategies is complicated by the fact that each strategy's effectiveness differs from one problem to the next: a strategy that works excellently for one type of problem may fail miserably on a different kind of problem. According to the SCADS model by Shrager and Siegler (1998), children solve this problem by gradually strengthening the association between the type of the problem solved and the strategy used after every correct answer. However, this model presupposes that children already know how to categorize problems in such a way that problems within the same category require the same strategies. Furthermore, the SCADS model presumes that learning is driven solely by whether or not the strategy produced the correct answer. This ignores the effort and time required to execute the strategy, and the mechanism is difficult to apply when performance feedback is continuous, as in economic decisions, rather than binary. Furthermore, even when those limitations are overcome the specific learning mechanism of the SCADS model appears to fail in some situations in which humans succeed (Lieder et al., 2014).

Our rational metareasoning model overcomes these limitations of the SCADS model. It could thus be used to model strategic development in domains that do not comply with the assumptions of the SCADS model. However, the applicability of our model to cognitive development remains to be evaluated. In this section we provide a proof of concept that our model can capture the developmental progression of children’s cognitive strategies in the domain of mental arithmetic. To do so, we simulate the development of children’s strategies for mental addition (Svenson & Sjöberg, 1983) according to rational metareasoning.

We recreated Shrager and Siegler's simulation of the development of children’s strategy use for single-digit addition problems in which both summands lie between 1 and 5 (Shrager & Siegler, 1998; Svenson & Sjöberg, 1983) with our strategy selection model. To make the model predictions as comparable as possible, we retained all of the assumptions that Shrager and Siegler made about children’s strategies. Concretely, we assumed that children use the following four strategies for mental addition:

1. *Retrieval*: retrieve the answer from memory.
2. *Sum*: First, use the fingers of one hand to count up to the first summand. Then use the fingers of the other hand to count up the second summand. Finally count the total number of raised fingers on either hand.
3. *Shortcut Sum*: After counting up to the first summand, continue counting upwards to the sum.
4. *Min*: Start counting upwards from the larger summand.

These four strategies differ in how many counting operations they require to solve any given problem. To account for the discovery of the Shortcut Sum strategy and the Min strategy, our metareasoning models start out with only the Retrieval strategy and the sum strategy. The Shortcut Sum strategy and the Min strategy are added after 90 and 95 trials respectively because this is how long it took children to discover those strategies in a study by Siegler and Jenkins (1989). To simulate reaction times, we assumed that each counting operation takes about half a second as indicated by the findings of Geary and Brown (1991). Following Shrager and Siegler (1998), errors were modeled by assuming that each counting step is incorrectly executed with probability $p_{\text{error}} = 0.04$. We generated the number of incorrectly executed steps by drawing from the binomial distribution $\text{Binomial}(\#\text{steps}, p_{error})$. The effect of each error was to either omit a counting operation, for example “3,3” instead of “3,4”, or to skip a number, for example “3,5” rather than “3,4”.

To model the Retrieval strategy, we modeled children’s memory for arithmetic facts by the associative memory model used in the SCADS model (Shrager & Siegler, 1998; Siegler & Shipley, 1995; Siegler & Shrager, 1984) with the same set of parameters.

This model characterizes memory for arithmetic facts by how strongly each possible answer a is associated with each problem $x+y$. The state of a child's long-term memory for arithmetic facts can therefore be described by a three-dimensional matrix $A(a,x,y)$ of associative strengths. For the most familiar addition problems whose first or second summand was 1 the associative strength was initialized with 0.05. For all other addition problems, the associative strengths were initialized by $1/(10 \cdot \#\text{values})$. Each time a strategy produced an answer the strength of the association between the answer and the pair of summands was increased by 0.06 if the answer was correct or by 0.03 when the answer was wrong. Each time the Retrieval strategy is used it samples a confidence criterion between 0 and 1 uniformly at random. The probability that a potential answer will be sampled is its associative strength divided by the sum of the associative strengths of all possible answers. If the associative strength of the sampled answer exceeds the confidence criterion, then the answer is reported. Otherwise the sampling process is repeated. If no answer's associative strength exceeded the confidence threshold after 10 attempts, then the Retrieval strategy fails to answer the question. The execution time of the Retrieval strategy was modeled as 0.5 seconds times the number of retrieved answers.

To apply our rational strategy selection learning model to mental addition, we have to specify how problems are represented, the form of the meta-level model, and children's prior knowledge about the performance of addition strategies. We assume that children represent the addition problem $x+y=?$ by three simple features

$$\mathbf{f}=(f_1,f_2,f_3)=\left(s_1,s_2,\max_{\mathrm{a}} A(a,x,y)\right),$$

where the third feature is the associative strength of the answer that is most strongly associated with the problem in memory. Since the feedback that children receive in mental addition is binary ("right" or "wrong"), the meta-level model is

$$P(R=1|\mathrm{f},S=i)=\frac{1}{1+\exp\left(b_i+\sum_{k=1}^{3}\alpha_{k,i}^{(R)}\cdot f_k\right)},$$

where the bias term b_i captures influences on the strategy's performance that are not captured by the features of the problem to be solved. We model children's prior knowledge about the performance of addition strategies by the model's prior on the bias weights. The simulations by Shrager and Siegler (1998) and Siegler and Shipley (1995) assumed that children initially know only the *Retrieval strategy* and the *Sum* strategy but have to discover the more efficient strategies on their own, since parents teach the *Sum* strategy first and memory retrieval is a domain general capacity that precedes knowledge of arithmetic. To capture these assumptions our simulations assume that children's prior expectation about the strategies' performance is positive for the *Sum* strategy ($P(b_2)=\mathcal{N}(5,1)$), neutral for the familiar *Retrieval strategy* ($P(b_2)=\mathcal{N}(0,1)$), but negative for the other strategies that are still unfamiliar ($P(b_3)=P(b_4)=P(b_5)=\mathcal{N}(-5,1)$). As in all previous applications of our model, the meta-level model uses Bayesian linear regression to predict each strategy's execution time from each problem's features. The relative cost of time was set such that finding the correct answer was worth 100 seconds. Since this corresponds to each child's subjective utility of being correct, this simulation assumed that the opportunity cost is known and does not have to be learned. To determine the predictions of our rational metareasoning model, we simulated the 200 virtual participants' choices of addition strategies across 100 blocks of 10 addition problems each. Addition problems were independently generated by randomly sampling the first and the second summand from two independent uniform distributions over their possible values, that is 1,2,3,4, or 5.

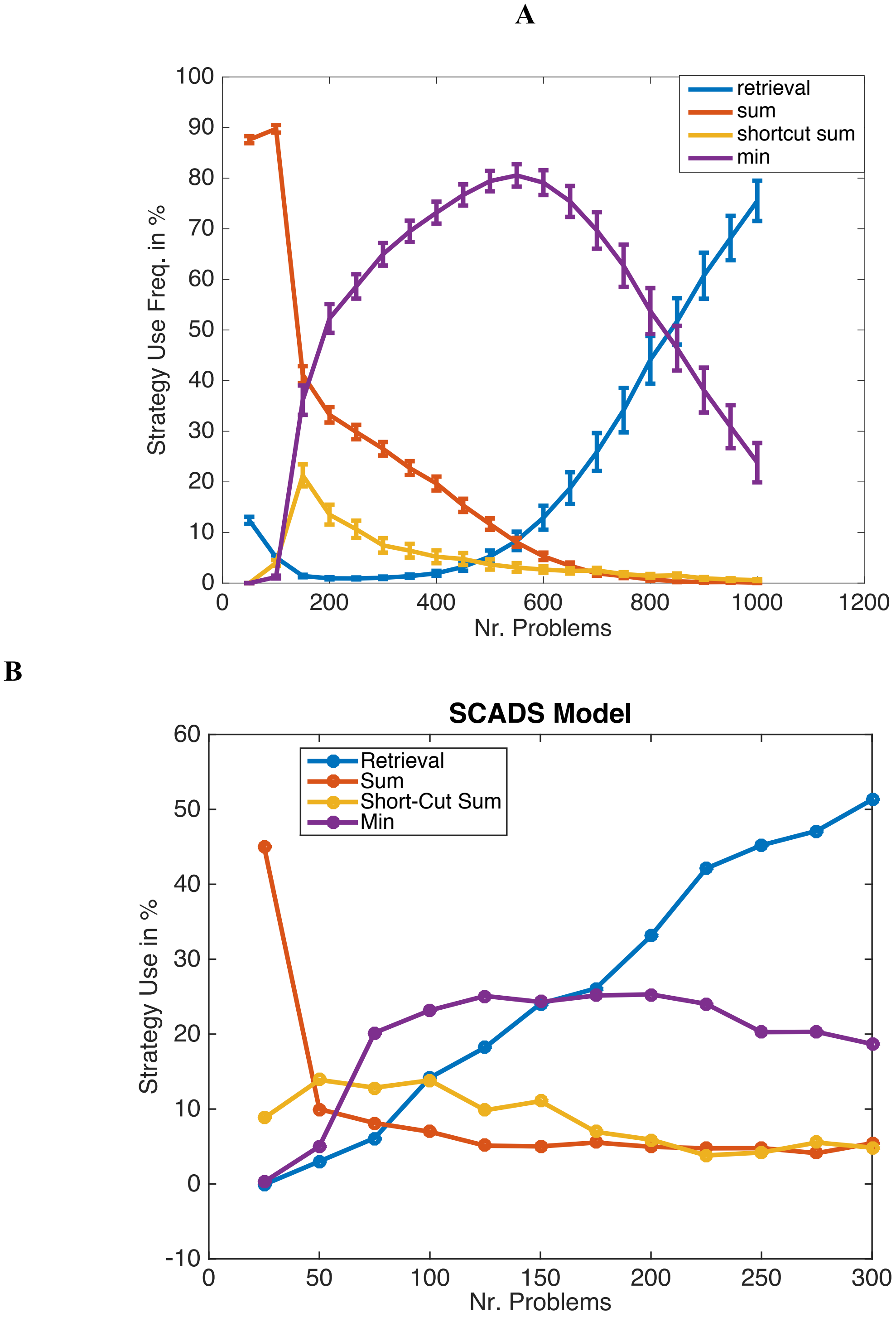


Figure 15: Comparison of models of how children learn to select arithmetic strategies. Error bars enclose 95% confidence intervals. (A) Predictions of rational metareasoning. (B) Predictions of the SCADS model according to Shrager and Siegler (1998).

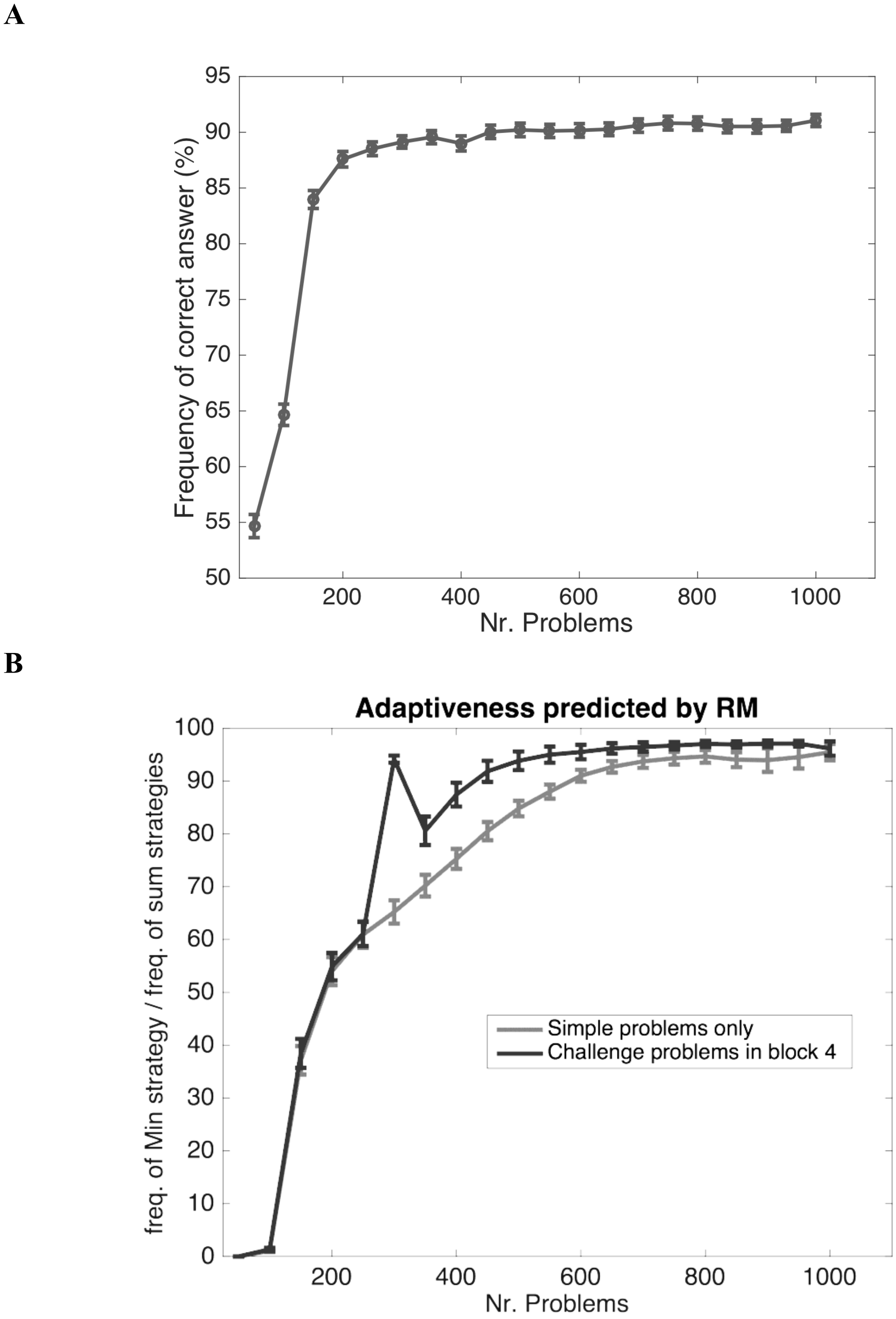


Figure 16: Learning curves of rational metareasoning simulations of children's strategic development in mental arithmetic. Error bars enclose 95% confidence intervals. (A) Gradual increase in performance predicted by rational metareasoning (RM). (B) Generalization of the Min strategy with versus without challenge problems.

The simulation results shown in Figure 15 suggest that our rational theory of strategy selection learning can capture the qualitative changes in children's use of addition strategies observed by Svenson and Sjöberg (1983): Our simulation captures the transient rise and fall of accurate but effortful addition strategies, the shift toward the more efficient *Min strategy*, and the eventual transition towards the predominant use of the *Retrieval strategy*. Comparing the predictions of our model with those of the SCADS model (Figure 15) suggests that both models capture the same developmental trends about equally well. Furthermore, like the SCADS model, rational metareasoning also captures the gradual increase in children's performance (see **Error! Reference source not found.**A) and the transfer from simple addition problems with summands ranging from 1 to 5 to more challenging addition problems with one addend above 20 and the other addend below 5. As shown in **Error! Reference source not found.**B, rational metareasoning predominantly selected the most accurate and the most efficient approach, namely the *Min strategy*, to solve the challenge problems even though it had never encountered any of those problems before.

Like the SCADS model, our model captures the increasingly adaptive strategy choices that children make: our model learned to use the Retrieval strategy more often for easy problems than for hard problems. This is adaptive because the Retrieval strategy is less accurate on hard problems because, due to past mistakes, hard problems are more strongly associated with wrong answers than easy problems. In our simulation, the correlation between a participant's average performance on a problem and the frequency with which they used the *Retrieval strategy* increased from $r(4998) = -0.11$ ($p < 0.001$) in the first 500 problems to $r(4998) = 0.28$ ($p < 0.001$) in problems 501 to 1000. In addition, our model learned to choose the *Min strategy* over less efficient and more error-prone addition strategies when the *Retrieval strategy* appeared inapplicable. Furthermore, the model learned to choose the *Min strategy* adaptively: The advantage of the *Min strategy* over alternative addition strategies increases with the sum and the difference between the addends. Across all simulated trials, the model's choice of the *Min strategy* was significantly correlated with the sum ($r(141609) = 0.30, p < 0.001$) and the absolute value of the difference between the addends ($r(141609) = 0.24,p < 0.001$). Furthermore, the correlation with the sum or the difference was stronger than the correlation with other factors such as the product ($r(141609) = 0.20,\ p < 0.001$). In addition, the model's choices of the *Min strategy* became more adaptive: Shortly after the discovery of the Min strategy (trials 100-150 to be precise) its use was less well predicted by the difference between the two summands ($r(4998) = 0.17, p < 0.001$) than by their product ($r(4998) = 0.32, p < 0.001$), but ten blocks later the difference between the two summands predicted the choice of the *Min strategy* ($r(4998) = 0.23, p < 0.001$) better than their product ($r(4998) = 0.09, p < 0.001$) as in Siegler and Shipley (1995).

As shown in **Error! Reference source not found.**B, the proportion of applications of the *Min strategy* out of all addition strategies increased steadily from 37.2% in the first 50 trials after its discovery towards 100%. The learning curve shows that the process by which the *Min strategy* is generalized from one problem on which it worked well to all other problems is gradual and takes more than 1000 examples. This is consistent with the empirical finding that children are slow to generalize the *Min strategy* to other problems upon its discovery. Siegler and Jenkins (1989) found that the generalization of the *Min strategy* proceeds much more rapidly when children who have recently discovered the *Min strategy* are posed challenge problems such as $4 + 25$. Shrager and Siegler (1998) modeled the experiment by Siegler and Jenkins (1989) by replacing the 50 simple problems presented after the discovery of the *Min strategy* by 50 challenge problems in which one of the addends is larger than 20 and the other addend is smaller than 5. We performed the equivalent simulation with our rational metareasoning model by replacing the 50 problems following the first five blocks by 50 challenge problems. As shown in **Error! Reference source not found.**B, feature-based strategy selection learning captures the empirical finding that challenge problems boost children's transfer of the *Min strategy* to challenging as well as simple problems (Siegler & Jenkins, 1989). To test if the observed differences were significant, we performed one t-test for each of the 20 simulated blocks of 50 problems with the Bonferroni-corrected significance level of $0.05/20 = 0.0025$. We found that the average adaptivity was not significantly different before the challenge problems (all $p \geq 0.20$) but became highly significant once the challenge problems were introduced ($p < 0.001$) and remained statistically significant until block 19 (all $p \leq 0.02$) after which the performance of both groups reached its asymptote (all $p \geq 0.50$).

To determine which components of our model were critical to capture the development of children's choice of addition strategies, we reran the simulation with the five lesioned metareasoning models. We found that exploration is necessary for strategic development, because without exploration the rational metareasoning model never discovered the Shortcut Sum strategy or the Min strategy, and it failed to switch to the Retrieval strategy even after it had plenty of experience to rely on (Supplementary Figure 10). Feature-based strategy selection was also critical, because the metareasoning model without features predicted that children would transition directly from the Sum Strategy to the Retrieval strategy without using the Shortcut Sum or the Min strategy in between (Supplementary Figure 11). This might be because the features are necessary to learn that the Retrieval strategy works only when the problem is familiar whereas the Min Strategy is superior for unfamiliar problems where one of the addends is small. Likewise, the lesioned metareasoning model that maximized accuracy regardless of time cost never discovered the Min strategy or the Shortcut Sum strategy but transitioned directly from the standard Sum strategy to memory retrieval (Supplementary Figure 12). Model-free metacognitive reinforcement learning of the VOC ($r = \text{reward-cost}$) predicted that the Sum strategy would fade much faster than it has been observed in children and failed to predict children's eventual transition to the Retrieval strategy (cf. Figure 14B vs. Supplementary Figure 13) Finally, model-free learning of the reward rates predicted an almost instantaneous shift to the Min strategy and also failed to predict the subsequent transition to the Retrieval strategy (see Supplementary Figure 14). These findings suggest that maintaining separate representations of execution time, opportunity cost, and expected reward enables faster learning and adaptation to changes in the strategies' performance or the reward rate.

In this section, we have demonstrated that rational metareasoning can explain several qualitative features of the shifts in children's choice of addition strategies. Most importantly, feature-based strategy selection learning formalizes the overlapping waves theory of cognitive development (Siegler, 1996) by a powerful, general learning mechanism. This suggests that our model should be able to capture similar phenomena in other domains of cognitive development as well. However, the change in children's strategy choices explained by our model is only one of three parts of strategic development, which also includes the discovery of new strategies and the change of existing strategies. To overcome this limitation, future work should combine our model of strategy selection learning with models of strategy discovery and strategy change. We will revisit this future direction in the General Discussion.

Feature-based strategy selection learning is more widely applicable than the basic SCADS model. Unlike the SCADS model our model can also learn from continuous feedback, as well as execution time or mental effort, and it does not presuppose that problems can be categorized appropriately. On the other hand, the SCADS model captures an important mechanism that is not yet included in our resource-rational account of strategic development: strategy discovery. Both mechanisms play an important role in strategic development. Therefore, our contributions are more complementary than they are in competition. Formalizing the computational mechanisms of strategy discovery and the formation of mental habits within the rational metareasoning framework is a promising direction for future research. To apply rational metareasoning to the strategy discovery problem, future research might combine learning to predict the VOC of individual computations from features of the current mental state with techniques from hierarchical reinforcement learning (Barto, Singh, & Chentanez, 2004; Barto & Mahadevan, 2003; Botvinick, Niv, & Barto, 2009; Sutton, Precup, & Singh, 1999).

General Discussion

How do we know when to think fast and when to think slow? Do we use our heuristics rationally or irrationally? How good are we at selecting the right strategy for the right problem? To answer these questions, we derived a rational solution to the strategy selection problem and evaluated it against human behavior and previous theories of strategy selection.

Our results support the conclusion that people gradually learn to use their cognitive strategies more rationally. According to our rational metareasoning model, these adaptive changes result from a rational metacognitive learning mechanism that builds a predictive model of each strategy's execution time and accuracy. Jointly, the experiments, simulations, and model comparisons reported in this article provided very strong evidence for all four components of our model: strategy selection based on an approximate cost-benefit analysis, feature-based metacognitive reinforcement learning, separate predictive models of accuracy and execution time, and the exploration of alternative strategies.

Our model's predictions captured the variability, contingency, and change of people's strategy choices in domains ranging from sorting to decision-making, and mental arithmetic as well as problem solving (see Supplementary Online Material). Our model provides a unifying explanation for a number of phenomena that were previously explained by different models. Overall, the dependence of people's strategy choices on task and context variables was consistent with a rational strategy selection mechanism that exploits the features of each problem to achieve an optimal cost-benefit tradeoff. Likewise, the change in people's strategy choices over time was consistent with rational learning of a predictive model of each strategy's performance and choosing strategies rationally with respect to the model learned so far. This learning mechanism simultaneously accounts for the developmental progression of children's arithmetic competence on a time scale of years and the adaptions of adults' decision strategies on a time scale of minutes. The remaining variability of people's strategy choices was consistent with the near-optimal exploration-exploitation tradeoff of Thompson sampling.

Critically, our new experiments and simulations showed that our model captures people's capacity to adapt to heterogeneous environments where each problem is unique and may require a different strategy than the previous one. Previous theories were unable to account for this adaptive flexibility but our rational account of strategy selection does. When we consider all of these phenomena jointly, our findings support the view that people choose cognitive strategies rationally subject to the constraints imposed by their finite time, limited information, and bounded cognitive resources. Its rational cost-benefit analysis allows our model to capture that people allocate their time and cognitive resources strategically so as to maximize their expected reward rate across multiple decisions rather than just their immediate reward on the current problem.

In addition, Experiments 2, 3, and 4 confirmed our model's prediction that resource-rationality increases with learning. In other words, people learn to make increasingly more rational use of their finite time and limited cognitive resources. Concretely, we found that people learn to think more when thinking is worthwhile and to think less when it is not. According to our theory, these adaptive changes result from metacognitive learning, and a person's experience is the primary limit on the rationality of their strategy choices.

Theoretical significance: implications for the debate about human rationality

Our theory reconciles the two poles of the debate about human rationality by suggesting that people gradually learn to make increasingly more rational use of fallible heuristics. Our emphasis on metacognitive learning provides a fresh alternative to previous accounts that viewed rationality as a fixed, static ideal, and irrationality as a pervasive trait. Instead, our theory suggests that we are constantly learning to think, learn, and decide more resource-rationally with respect to the problems, rewards, and costs we experience. Hence, if we engage seriously with the environments we want to master, then metacognitive learning should propel us towards bounded rationality as we learn to choose the strategies that achieve the best possible cost-benefit tradeoff. Thus, although we might never reach the ideals of (bounded) rationality, we can become a little more resource-rational every time we use a cognitive strategy. Whether these improvements depend on deliberate reflection is an interesting question for future research.

The strategy selection problem is a critical missing piece in the puzzle of what it means to be boundedly rational. Our proposal for a rational solution to the strategy selection problem might therefore be an important step towards a unifying theory of bounded rationality. Indeed, recent work suggests that rationally choosing among a *small* number of cognitive strategies is optimal for bounded agents (Milli, Lieder, & Griffiths, 2017). Our model solves the riddle how a bounded agent can possibly optimize the use of its limited resources by investing some of them into solving the computationally intractable and potentially recursive problem of optimizing the use of its limited resources. We have proposed that the mind side-steps the computational complexity and infinite regress of this problem by *learning*—rather than computing—the value of investing time and cognitive resources into one strategy versus another. We show that good strategies can be selected very efficiently once an approximation to the value of computation has been learned and that the learning process can be implemented very efficiently as well (see Figure 1). Despite its simplicity this mechanism can adaptively choose between complex and extremely time- and resource-consuming strategies. It may thereby enable the mind to save a substantial amount of cognitive resources and find good approximate solutions to intractable problems. Our model can therefore be used to complete dual-process theories of bounded rationality (Evans & Stanovich, 2013; Evans, 2003; Kahneman, 2011) by a rational, yet tractable, mechanism for determining when to employ which system. Our strategy selection mechanism could be integrated into dual-process theories to predict exactly when people think fast and when they think slow. Likewise, our mechanism could also be integrated into adaptive toolbox theories of bounded rationality (Todd & Brighton, 2015; Todd & Gigerenzer, 2012) to predict exactly which heuristic people will use in a given situation. This line of research would lead to mathematically precise, falsifiable theories of bounded rationality that could be quantitatively evaluated against empirical data and each other.

Our perspective emphasizes the importance of metacognitive values for human rationality. This emphasis is consistent with the view that individual differences in rationality reflect people's dispositions towards different cognitive styles ("the reflective mind") rather than their cognitive abilities per se ("the algorithmic mind", Stanovich, 2011). Our theory suggests that the disposition towards rational versus heuristic thinking is not fixed and innate but malleable and learned from experience. Yet, our theory also suggests that a person's propensity for rational thinking can be highly situational because the mind estimates the value of deliberation from contextual features.

Future directions

Future work should extend the proposed model to capture additional aspects of human cognition. One such extension could be a more realistic model of the cost of strategy execution which captures that some strategies are more effortful than others. This could be achieved by modeling how much cognitive resources, such as working memory, each strategy consumes at each point in time. With this extension, the total cost of executing a strategy could be derived by adding up the opportunity costs of its consumed resources over the time course of its execution.

While our model comparisons show that strategy selection learning requires some form of exploration, it is silent about how this exploration is accomplished. The Thompson sampling mechanism evaluated here is one of the best solutions to the exploration-exploitation tradeoff known to date (Chapelle & Li, 2011; Kaufmann, Korda, & Munos, 2012), but many alternative exploration mechanisms have been proposed in the reinforcement learning literature. These proposals range from simple mechanisms like epsilon-greedy action selection and the soft-max decision rule (Sutton & Barto, 1998) to more sophisticated mechanisms including upper-confidence bound algorithms (Auer, 2002) and other exploration bonuses (Brafman & Tennenholtz, 2002). At this point, each of these algorithms is a viable hypothesis about human strategy selection, and designing experiments to test them is an important direction for future research.

While our simulations and model comparisons favored learning separate predictive models of execution time and accuracy over learning the VOC directly, this advantage might reflect specific, auxiliary assumptions of our model. A more definitive answer will require experiments that systematically disambiguate these two learning mechanisms. Based on how model-free and model-based control over behavior are usually disambiguated (Dickinson, 1985), strategy selection experiments that devaluate speed or accuracy (but not both) after people have learned to achieve the optimal speed-accuracy tradeoff might be a fruitful direction for future research.

Since our model is agnostic about the set of strategies people choose from, future work should determine which strategies are available to people. This could be done by comparing rational metareasoning models with different sets of strategies using Bayesian model selection (Scheibehenne, Rieskamp, & Wagenmakers, 2013).

People's decision mechanisms likely include strategies with continuous parameters, such as sequential sampling models with decision thresholds and attentional biases (Smith & Ratcliff, 2004), satisficing strategies with aspiration levels (Simon, 1955), and simulation-based decision mechanisms that can perform varying numbers of simulations (e.g., Lieder et al., 2014). Furthermore, the proposed process model only learns about a small subset of all possible cognitive strategies. To select among all possible sequences of elementary information processing operations, our process model has to be extended to learning the VOC of individual computations instead of only learning the VOC of complete strategies that always generate an action yielding reward. Current work is extending the proposed model to overcome these limitations (Krueger, Lieder, & Griffiths, 2017; Lieder, Krueger, & Griffiths, 2017; Lieder, Shenhav, Musslick, & Griffiths, 2017).

To capture people's ability to plan sequences of cognitive operations, future work might add predictive models for features of the agent's future internal states alongside the predictive models of the expected reward and execution time. This extension would correspond to learning option models (Sutton et al., 1999)—a form of model-based hierarchical reinforcement learning (Barto & Mahadevan, 2003; Sutton et al., 1999) that holds promise for explaining the complex hierarchical structure of human behavior (Botvinick & Weinstein, 2014). Both extensions could be combined with ideas from hierarchical reinforcement learning to capture how people discover novel, more effective strategies by flexibly combining elementary operations with partial.

The third major limitation of the current model is that it presupposes domain-specific problem features. A complete account of strategy selection would have to specify where those representations come from. To provide such an account, our model could be implemented as a hierarchical neural network with several layers in-between the perceptual input and the representation of the features as illustrated in Figure 1. In such a network the features could emerge from the same error-driven learning mechanism used to learn the weights between the feature layer and the layers representing the network's predictions (cf. Mnih et al., 2015).

Future experiments might also investigate whether the proposed feature-based strategy-selection mechanism coexists with a more basic, automatic strategy selection mechanisms based on context-free RL. If so, then our framework could be used to model the arbitration between them as rational meta-strategy-selection.

One important open theoretical question is under which, if any, conditions the proposed strategy selection mechanism is boundedly optimal (Russell & Subramanian, 1995). While it is possible to prove the optimality of a program for a particular computational architecture, such proofs have yet to be attempted for computational models of the human mind.

In addition to its contributions to the debate about human rationality and its utility for future basic research, our model of strategy selection learning might also have potential practical applications in education and cognitive training. In terms of education, our model could be used to optimize the problem sets used to teach students when to use which approach—for instance in mathematical problem solving or high school algebra. In terms of cognitive training, our model could be used to investigate which training regimens increase cognitive flexibility by promoting adaptive strategy selection. According to our theory, people's ability to (learn to) represent problems by general features that are predictive of the differential efficacy of alternative strategies would be a critical prerequisite for such training to succeed.

In conclusion, our findings paint an optimistic picture of the human mind by highlighting metacognitive learning and the resulting cognitive growth. This perspective highlights that our rationality is not fixed but malleable and constantly improving. We hope that specifying what people's metacognitive learning mechanisms might be, our model will give us a handle on how to leverage them to promote cognitive growth.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Taylor & Francis.

Ariely, D. (2009). *Predictably irrational*. New York: Harper Collins.

Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In J. Triesch & T. Jebara (Eds.), *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)* (pp. 112–119). San Diego, CA: UCSD Institute for Neural Computation.

Barto, A., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(4), 341–379. doi:10.1023/a:1025696116075

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*(3), 439–449. doi:10.1002/for.3980050302

Bjorklund, D. F., & Douglas, R. N. (1997). The development of memory strategies. In N. Cowan, and C. Hulme (Eds.), *The Development of Memory in Childhood* (pp. 201–246). Hove, East Sussex, UK: Psychology Press.

Botvinick, M., Niv, Y., & Barto, A. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. doi:10.1016/j.cognition.2008.08.011

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655). doi:10.1098/rstb.2013.0480

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences*, *19*(11), 700-710. doi:10.1016/j.tics.2015.08.013.

Braver, T. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi:10.1016/j.tics.2011.12.010.

Bröder, A. (2003). Decision making with the“ adaptive toolbox”: Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 611-625. doi:10.1037/0278-7393.29.4.611

Daw, N., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. doi:10.1038/nn1560

Dolan, R., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. doi:10.1038/nn1560

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193. doi:10.1037/h0044139

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*(4), 912–931. doi:10.1037/0033-295X.112.4.912

Evans, J. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. doi:10.1016/j.tics.2003.08.012

Evans, J. S., & Stanovich, K. E. (2013). Dual-process theories of higher cognition. *Perspectives on Psychological Science*, *8*(3), 223–241. doi:10.1177/1745691612460685

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. doi:10.1037/0003-066x.34.10.906

Fum, D., & Del Missier, F. (2001). Adaptive selection of problem solving strategies. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Meeting of the Cognitive Science Society* (pp. 313–318). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gigerenzer, G. (2008a). *Rationality for mortals: How people cope with uncertainty*. New York: Oxford University Press.

Gigerenzer, G. (2008b). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29. doi:10.1111/j.1745-6916.2008.00058.x.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. doi:10.1146/annurev-psych-120709-145346

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

Gigerenzer, G., Todd, P. M., & A.B.C. Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University 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. doi:10.1111/tops.12142.

Gunzelmann, G., & Anderson, J. R. (2001). An ACT-R model of the evolution of strategy use and problem difficulty. In E. M. Altmann, A. Cleermans, C. D. Schunn, & W. D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling* (pp. 109–114). Mahwah, NJ: Lawrence Erlbaum Associates.

Gunzelmann, G., & Anderson, J. R. (2003). Problem solving: Increased planning with practice. *Cognitive Systems Research*, *4*(1), 57–76. doi:10.1016/s1389-0417(02)00073-6

Hay, N., Russell, S., Tolpin, D., & Shimony, S. (2012). Selecting computations: Theory and applications. In N. de Freitas & K. Murphy (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Eighth Conference*. Corvallis, Oregon: AUAI Press. Retrieved from http://arxiv.org/abs/1207.5879.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, *31*(4), 395–414. doi:10.1287/mnsc.31.4.395

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi:10.2307/1914185

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.2307/2291091.

Kawaguchi, K., Kaelbling, L. P., & Lozano-Pérez, T. (2015). Bayesian Optimization with Exponential Convergence. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 2809–2817). Curran Associates, Inc.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Computational Biology*, *7*(5), e1002055. doi:10.1371/journal.pcbi.1002055.

Knuth, D. E. (1998). *The art of computer programming: sorting and searching* (Vol. 3). Boston: Pearson Education.

Kunz, S. (2009). *The Bayesian linear model with unknown variance*. Technical Report, University of Zurich.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*(6), 661–79. doi:10.1017/S0140525X12003196

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. doi:10.1111/tops.12086

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting 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 P. Bartlett, F. C. N. Pereira, L. Bottou, C. J. C. Burges, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*.

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 Meeting 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. In Z. Ghahramani, M. Welling, K. Q. Weinberger, C. Cortes, & N. D. Lawrence (Eds.), *Advances in Neural Information Processing Systems 27*.

Lindley, D. V, & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(1), 1–41. doi: 10.2307/2985048

Maass, W. (2000). On the computational power of winner-take-all. *Neural Comput.*, *12*(11), 2519–2535. doi:10.1162/089976600300014827

Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Boston: Houghton Mifflin Harcourt.

Marewski, J. N., & Link, D. (2014). Strategy selection: An introduction to the modeling challenge. Wiley Interdisciplinary Reviews: Cognitive Science, 5(1), 39–59. doi:10.1002/wcs.1265

Marewski, J., & Schooler, L. (2011). Cognitive niches: an ecological model of strategy selection. *Psychological Review*, *118*(3), 393–437. doi:10.1037/a0024143

Marr, D. (1983). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Company.

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, *13*, 2069–2106. Retrieved from http://dl.acm.org/citation.cfm?id=2188385.2343711

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi:10.1038/nature14236

Niv, Y., Daw, N., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. doi:10.1007/s00213-006-0502-4

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin*, *92*(2), 382. doi:10.1037/0033-2909.92.2.382

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534-552. doi:10.1037/0278-7393.14.3.534

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

Penny, W., & Ridgway, G. (2013). Efficient posterior probability mapping using Savage-Dickey ratios. *PLoS ONE*, *8*(3), e59655. doi: 10.1371/journal.pone.0059655

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236. doi:10.1037/0096-3445.135.2.207

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Articial Intelligence Research*, (2), 575–609. Retrieved from http://dl.acm.org/citation.cfm?id=1622826.1622844

Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, *49*(1-3), 361–395. doi:10.1016/0004-3702(91)90015-c

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*(5), 405–410. doi:10.1111/1467-9280.00076

Siegler, R. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275. doi:10.1037/0096-3445.117.3.258

Siegler, R., & Jenkins, E. A. (1989). *How children discover new strategies*. New York: Psychology Press.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford, UK: Oxford University Press.

Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Sciences*, *3*(11), 430–435. doi:10.1016/s1364-6613(99)01372-8

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Graeme (Eds.), *Developing Cognitive Competence: New approaches to process modeling* (pp. 31–76). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do. In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 229–293). Hillsdale, NJ: Larence Erlbaum Associates.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. doi:10.1037/h0042769

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. doi:10.2307/1884852

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

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. doi:10.1016/j.tins.2004.01.006

Stanovich, K. (2011). *Rationality and the reflective mind*. Oxford: Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*(5), 645. doi:10.1017/S0140525X00003435

Sutherland, S. (2013). *Irrationality: The enemy within*. London: Pinter & Martin Ltd.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1-2), 181–211. doi:10.1016/s0004-3702(99)00052-1

Svenson, O., & Sjöberg, K. (1983). Evolution of cognitive processes for solving simple additions during the first three school years. *Scandinavian Journal of Psychology*, *24*(1), 117–124. doi:10.1111/j.1467-9450.1983.tb00483.x

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 285–294. doi:10.2307/2332286

Todd, P., & Brighton, H. (2015). Building the theory of ecological rationality. *Minds and Machines*, 1–22. doi:10.1007/s11023-015-9371-0

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. New York: Oxford University Press.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168. doi:10.1080/13546783.2013.844729

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281. doi:10.1037/h0032955

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293. doi:10.1037/0033-295X.90.4.293

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. doi:10.1111/cogs.12101

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. doi:10.1080/14640746808400161

Watkins, C. F., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292. doi:10.1007/bf00992698

Appendix A

1. SSL and RELACS

According to the SSL model (Rieskamp & Otto, 2006) the probability that strategy i will be chosen ($P(S_t = i)$) in trial t is proportional to its reward expectancy q_t:

$$P(S_t = i) \propto q_t(i),$$

where $q_t(k)$ is the sum of the rewards obtained when strategy k was chosen prior to trial t plus the initial reward expectancy

$$q_0(k) = r_{max} \cdot w \cdot b_k,$$

where r_{max} is the highest possible reward, w is the strength of the initial reward expectancy, and $b_1, \cdots, b_N \in [0; 1]$ are the agent's initial relative reward expectancies for strategies $1, \cdots, N$ and sum to one.

The RELACS model (Erev & Barron, 2005) chooses strategies according to their recency-weighted average payoffs

$$w_{t+1}(k) = \begin{cases} \alpha \cdot r_t + (1-\alpha) \cdot w_t(k) & \text{if } S_t = k \\ w_t & \text{else} \end{cases}$$

$$P(S_t = k) \propto e^{\lambda \cdot \frac{w_t(k)}{V_t}},$$

where the parameters α and λ determine the agent's learning rate and decision noise respectively, and V_t is the agent's current estimate of the payoff variability.

2. Conjugate update equations for the posterior distribution of a Gaussian likelihood and a Gaussian prior

The prior on the reward rate is a normal distribution and the likelihood of the ratio of observed total reward over total time is a standard normal distribution, that is

$$P(\bar{r}) = \mathcal{N}(1,1),$$

$$P\left(\frac{r_{\text{total}}}{t_{\text{total}}}\middle|\bar{r}\right) = \mathcal{N}\left(\bar{r}, \frac{t_{\text{total}}}{60\text{sec}}\right).$$

Consequently, the posterior distribution of the reward rate is

$$P(\bar{r}|r_{\text{total}}, t_{\text{total}}) = \mathcal{N}(\mu_{\text{post}}, \tau_{\text{post}}),$$

with

$$\tau_{\text{post}} = \tau_{\text{prior}} + \tau_{\text{likelihood}} = 1 + \frac{t_{\text{total}}}{60\text{sec}}$$

$$\mu_{\text{post}} = \frac{\tau_{\text{prior}} \cdot \mu_{\text{prior}} + \tau_{\text{likelihood}} \cdot \frac{r_{\text{total}}}{t_{\text{total}}}}{\tau_{\text{prior}} + \tau_{\text{likelihood}}} = \frac{1 + \frac{r_{\text{total}}}{60\text{sec}}}{1 + \frac{t_{\text{total}}}{60\text{sec}}}.$$

3. Bayesian Regression

For continuous outcomes (i.e., execution time and reward) we performed exact Bayesian inference in a linear regression model (Kunz, 2009; Lindley & Smith, 1972). For binary outcomes (i.e., correct vs. incorrect) we use Bayesian logistic regression with the Laplace approximation (Lieder & Griffiths, 2015; Lieder et al., 2014). This approach learns a probability distribution over the amount of time that will pass and the amount of reward that will be obtained.

The Bayesian linear regression (Kunz, 2009; Lindley & Smith, 1972) model for the execution time T was defined as

$$P\left(T\middle|\mathbf{f}, s, w^{(T)}, \sigma_T^2\right) = \mathcal{N}\left(\mu = \sum_i w_{k,s}^{(T)} \cdot f_i \,, \sigma_T^2\right),$$

$$P\left(w_{:,s}^{(T)}\right) = \mathcal{N}(\mu = 0, \Sigma = \mathrm{Id}),$$

$$P(\sigma_T^2) = \mathrm{InvGamma}(\alpha_0, \beta_0),$$

where $\mathcal{N}$ stands for the normal distribution, InvGamma stands for the inverse gamma distribution, $w_{:,s}^{(T)}$ is the vector of the the weights of all features on the expected execution time of strategy s, and Id stands for the identity matrix. Given observed rewards $\boldsymbol{r}^{(\mathbf{1},\cdots,\boldsymbol{t})} = (r_1, \cdots, r_t)$ in trials $1, \cdots, t$ when the strategy was applied to a problem with features $\boldsymbol{f}^{(1,\cdots,t)} = \left(\boldsymbol{f}^{(1)}, \cdots, \boldsymbol{f}^{(t)}\right)$, i.e. $P\left(\alpha^{(s)}|\boldsymbol{r}, \boldsymbol{f}^{(1,\cdots,t)}\right)$ the model's prior distributions on the regression coefficient and the variance were updated to the respective posterior distributions $P\left(\alpha^{(T)}|\boldsymbol{r}^{(1,\cdots,t)}, \boldsymbol{f}^{(1,\cdots,t)}\right)$ and $P\left(\sigma_T^2|\boldsymbol{r}^{(1,\cdots,t)}, \boldsymbol{f}^{(1,\cdots,t)}\right)$. Since the priors are conjugate to the likelihood function, the posterior distributions are in the same family as the prior distributions and their parameters can be computed by the standard update equations for the normal-normal and normal-gamma models (Kunz, 2009; Lindley & Smith, 1972). When the reward was continuous, then the same model was used for learning to predict the reward. But if the reward was binary then we used Bayesian logistic regression with the Laplace approximation (see Section 3).

The model's priors on the error variance and the precision of the prior on regression coefficients were set to convey weak domain knowledge. In the sorting simulations, the prior expectation on the variance of the noise was 10 for the execution time in seconds ($\alpha_0 = 1, \beta_0 = 10$) and 0.1 for the binary reward ($\alpha_0 = 10, \beta_0 = 1$), and the standard deviation of the prior on the regression coefficients was 10 for the execution time and 1 for the binary score. These priors reflect that the execution times in this simulation were one to two orders of magnitude larger than the rewards.

In the simulations of the decision-making experiments by Payne et al. (1988), the prior expectation of the variance in the execution time was 1 ($\alpha_0 = \beta_0 = 1$), and the variance of the prior on the coefficients predicting the execution time was 1 as well. Since the relative reward was confined to the interval $[-1,1]$ the prior expectation of its error variance was 0.1 ($\alpha_0 = 1, \beta_0 = 0.1$); the precision of the prior on the regression coefficients was 1.

The simulations of the Mouselab experiments assumed a time cost of $7/h at a rate of 1 computation/sec. The prior on the reward rate corresponded to 1 minute's worth of experience in an environment with a reward rate of $7/h. The prior distributions on the strategies expected rewards and execution times were a normal distribution with mean zero and precision 0.1. The priors on the error variances of execution time and expected reward were $\mathrm{Gamma}(1,1)$.

In the simulations of the Rieskamp experiments the precision of the Gaussian prior on the coefficients of the reward model and the execution time model were estimated according to the maximum likelihood method. The prior on the error variance of the score model was $\mathrm{Gamma}(1,0.1)$ and the prior on the error variance of the execution time was $\mathrm{Gamma}(1,1)$.

In the simulations of mental arithmetic, the variance of the prior on the regression coefficients was 1 for both the execution time model and the model of accuracy, because the score was binary and single-digit addition takes only a few seconds. The prior on the error variance of the execution time was $\mathrm{Gamma}(1{,}1)$ because the execution time variability of addition strategies is in the order of seconds.

4. Laplace Approximation to Bayesian Logistic Regression

When the reward is binary (e.g., correct versus incorrect) rather than continuous, then linear regression would be ill-suited to predict it. Hence, in this case our model uses Bayesian logistic regression to predict that probability that the response will be correct ($R = 1$). According the Bayesian logistic regression model, the probability that a strategy s will generate a reward is given by

$$P(R = 1|s,\mathbf{f}, \alpha) = \frac{1}{1 + \exp\left(-\sum_k w_{k,s}^{(R)} \cdot f_k\right)},$$

$$P\left(\alpha^{(s)}\right) = \mathcal{N}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 0.01 \cdot \mathbf{I})$$

The posterior distribution on the regression coefficients $w_{:,s}^{(R)}$ for the expected reward of strategy s given observed rewards $\boldsymbol{r}^{(\mathbf{1},\cdots,\boldsymbol{t})} = (r_1, \cdots, r_t)$ in trials $1, \cdots, t$ when the strategy was applied to a problem with features $\boldsymbol{f}^{(1,\cdots,t)} = \left(\boldsymbol{f}^{(1)}, \cdots, \boldsymbol{f}^{(t)}\right)$, i.e. $P\left(w_{:,s}^{(R)}|\boldsymbol{r}, \boldsymbol{f}^{(1,\cdots,t)}\right)$, does no longer have a simply analytic solution. Therefore, we approximate by a normal distribution whose mean is the mode of the posterior distribution and whose precision matrix is the negative Hessian (which is the matrix of second partial derivatives) of the log-posterior at its mode:

$$\begin{aligned} P\left(w_{:,s}^{(R)}|\boldsymbol{r}^{(1,\cdots,t)}, \boldsymbol{f}^{(1,\cdots,t)}\right) &\approx Q\left(w_{:,s}^{(R)};\ \boldsymbol{r}^{(1,\cdots,t)}, \boldsymbol{f}^{(\mathbf{1},\cdots,\boldsymbol{t})}\right) \\ &= \mathcal{N}(\mu = w_{\max}, \Sigma^{-1} = -H(\alpha_{\max})), \\ w_{\max} &= \arg\max_{\alpha} p\left(w_{k,s}^{(R)} = w|\boldsymbol{r}, \boldsymbol{f}\right) \\ H_{i,j} &= \frac{\partial^2 \log p\left(w_{:,s}^{(R)}\middle|\boldsymbol{r}, \boldsymbol{f}\right)}{\partial\, \alpha_i^{(s)} \partial\, \alpha_j^{(s)}}. \end{aligned}$$

This is known as the Laplace approximation. It can be derived as a second-order Taylor series expansion of the log-posterior. The posterior mode was determined by numerical optimization using the function *fminunc* from the Matlab 2014b optimization toolbox and the gradients and Hessian were computed analytically.

5. Feature Selection by Bayesian Model Selection

To model how people discover which features are relevant for predicting a strategy's execution time or reward, our model includes a feature selection mechanism. According to our model, features are selected by Bayesian model selection (Kass & Raftery, 1995). Concretely, we consider one model for each possible subset of the features and determine the model with the highest posterior probability given the observations. To efficiently compute Bayes factors, we exploit that all models are nested within the full model that includes all of the features by computing Savage-Dickey ratios (Penny & Ridgway, 2013).