Journal of Experimental Psychology:
Learning, Memory, and Cognition
2015, Vol. 41, No. 1, 193–208

0278-7393/15/$12.00 http://dx.doi.org/10.1037/a0038118

Sell in May and Go Away? Learning and Risk Taking in Nonmonotonic Decision Problems

Renato Frey
Max Planck Institute for Human Development

Jörg Rieskamp
University of Basel

Ralph Hertwig
Max Planck Institute for Human Development

In nonmonotonic decision problems, the magnitude of outcomes can both increase and decrease over time depending on the state of the decision problem. These increases and decreases may occur repeatedly and result in a variety of possible outcome distributions. In many previously investigated sequential decision problems, in contrast, outcomes (or the probabilities of obtaining specific outcomes) change monotonically in 1 direction. To investigate how and to what extent people learn in nonmonotonic decision problems, we developed a new task, the Sequential Investment Task (SIT), in which people sequentially decide whether or not to sell shares at several selling points over the course of virtual days. Across trials, they can learn which selling point yields the highest payoff in a specific market. The results of 2 experiments suggest that a reinforcement-learning model generally describes participants' learning processes best. Learning largely depends on an interaction of the complexity of the stochastic process that generates the outcome distribution (i.e., whether the peak selling point is early or late in the selling period and whether there are single or multiple payoff maxima) and the amount of feedback that is available for learning. Although the risk profile in nonmonotonic decision problems renders exploration relatively safe, a clear gap persisted between the choices of people receiving partial feedback (thus facing an exploration–exploitation trade-off) and those of people receiving full feedback: Only the choices of the latter consistently approximated the peak selling points.

Keywords: sequential decision making, risk taking, reinforcement learning, computational modeling, exploration–exploitation trade-off

In various domains of life, people face challenging sequential decision problems that involve risks. In the stock market, for example, investors sequentially decide whether to buy or sell particular stock and, in so doing, face the problem of "market timing" (Benartzi & Thaler, 2007); that is, finding the best time to buy or sell. Some investors rely on the "sell in May and go away" strategy, believing that stock markets begin a downward trend in May and resume growth in October. In fact, whereas data from the years 1970 to 1998 support this belief (Bouman & Jacobsen, 2002), more recent analyses suggest that investors following this strategy have missed out on about 50% of returns over the past 10 years (Caldwell, 2013). Whether or not learning about features of the stochastic processes involved in the stock market can increase performance is disputed, but investment decisions indeed seem to be influenced by recent market trends. For example, in bear markets (when the price trend is down), investors are pessimistic and the tendency to sell shares is strongest (De Bondt, 1993; Schachter, Ouellette, Whittle, & Gerin, 1987), causing them to "buy high but sell low" (Benartzi & Thaler, 2007). One possible explanation for this observation is that investors who have sold their shares lack feedback and the experience of a potential trend reversal, which might have influenced their future selling behavior.

In this article, we investigate how people learn to make sequential decisions in environments such as the stock market, in which outcomes can fluctuate over time. In general, sequential decision problems are challenging, because people are able to learn about the possible outcomes only through experience. Identifying the state of the sequential decision problem that yields the highest outcome may not be trivial for at least two reasons. First, people may face a trade-off between learning (exploration) and deciding based on the acquired knowledge (exploitation). Second, sequential decision problems are difficult because certain outcomes can

This article was published Online First November 24, 2014.

Renato Frey, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; Jörg Rieskamp, Department of Psychology, University of Basel, Switzerland; Ralph Hertwig, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

We thank the Swiss National Science Foundation for a grant (CRSII1_136227) to the second and third authors; Carmen Kaiser for collecting the data; Laura Wiles and Susannah Goss for editing the manuscript; and Adrian Bruhin, Tomás Lejarraga, Andreas Pedroni, and Timothy J. Pleskac for their helpful comments, which have substantially improved this article.

Correspondence concerning this article should be addressed to Renato Frey, Max Planck Institute for Human Development, Center for Adaptive Rationality, Lentzeallee 94, 14195 Berlin, Germany. E-mail: frey@mpib-berlin.mpg.de

be obtained only by a sequence of interdependent decisions. For example, people may have to accept momentarily declining trends to reach subsequent states that in turn yield higher outcomes. People can use different learning strategies in the context of sequential decision problems, and which strategy they eventually adopt may depend on both the structure of the environment and the information available for learning about the environment.

Given the challenges of sequential decision problems, it is not surprising that they have attracted attention from a variety of research fields. Statisticians, mathematicians, and computer scientists have explored the optimal solutions to sequential decision problems (Ferguson, 1989) and how well different learning models can solve the respective tasks (Sutton & Barto, 1998). In psychology, sequential decision problems have often been employed to gauge people's risk attitudes and to predict real-life risk taking (e.g., Hoffrage, Weber, Hertwig, & Chase, 2003; Pleskac, 2008; Wallsten, Pleskac, & Lejuez, 2005). Recent work in neuroscience has also explored interindividual differences in healthy and clinical populations (Chiu, Lohrenz, & Montague, 2008; Frank, Doll, Oas-Terpstra, & Moreno, 2009).

Different Environments of Sequential Decision Problems

In a sequential decision problem, a specific stochastic process governs how outcomes or the probabilities of obtaining specific outcomes change as a function of the decision problem's sequential state. This stochastic process can adopt a variety of characteristics, resulting in different environments of sequential decision problems. A commonly investigated environment is one in which the outcomes or the probabilities of obtaining outcomes change *monotonically* from state to state until the sequential decision problem ends. An example of such a "monotonic decision problem" is the game blackjack, in which each draw of a card monotonically increases the current score but, at the same time, also monotonically increases the risk of going bust (i.e., exceeding a score of 21). However, many other sequential decision problems in real life—including the stock market, the progress of a chronic disease, and the prices of products such as airline tickets—do not obey such a monotonicity: Outcomes can increase and decrease over time, and trends may replicate themselves. Furthermore, there is not necessarily the risk of a sudden and complete loss if too many states of the decision problem are explored. This article makes the novel contribution of investigating people's learning strategies in such "nonmonotonic decision problems." As monotonic and nonmonotonic decision problems naturally share certain similarities, we first review challenges to learning in monotonic decision problems before turning to the approach we take to study nonmonotonic decision problems.

Monotonic Decision Problems

At each state of a monotonic decision problem, people face a choice between (a) proceeding to the next state, which may increase the current payoff by some amount or may terminate the problem (and often results in the total loss of the accumulated payoff), and (b) stopping at the current state of the problem and cashing in the accumulated payoff. A prominent paradigm that implements such sequential decisions is the Balloon Analogue Risk Task (BART; Lejuez et al., 2002). In the BART, people pump up virtual balloons over multiple trials. Each pump linearly increases the payoff by a fixed amount. But with every pump, the probability also increases that the balloon will burst and the money accrued will be lost. This risk increases in a monotonic and nonstationary way, such that the closer the number of pumps is to a balloon's maximum capacity (which is initially unknown and has to be explored), the more likely the risk of an explosion.

Challenges to learning. First, learning in a sequential decision problem may be difficult depending on the complexity of the underlying stochastic process. In the BART, the basic features of this process can be inferred relatively easily from the properties of real balloons: It is evident a priori that there is only one maximum, meaning that the probability of an explosion increases steadily up to a certain state and that exceeding this state terminates the current trial. The challenge to learning is thus to find the state with the highest expected outcome, while taking into account the key properties of the underlying stochastic process, such as how quickly the probability of an explosion grows. Another challenge to learning is the required trade-off between exploration and exploitation. In the BART, for example, people receive only partial feedback. In particular, they do not learn how many additional times they could have pumped once they decide to stop pumping. Thus, there is an asymmetry in error correction (Denrell, 2007; Denrell & March, 2001; Einav, 2005): The error of pumping too often results in an explosion, and this experience can help to reduce the number of pumps in the following trial. The error of pumping too little, in contrast, does not lead to distinct feedback from the environment. The second error—pumping to little—is less costly than the total loss due to an explosion. Nevertheless, people may use a certain number of trials for exploration (and pump more than they believe the balloon's maximum capacity to be) before they start to exploit their knowledge and pump up to the state that they believe to yield the highest expected outcome. Partial feedback in the BART may be one of the reasons why people generally appear relatively risk averse and pump too little to maximize their earnings (Lejuez et al., 2002).

Learning strategies. Various computational models have been proposed to describe learning in sequential decision problems such as the BART. The target model (Wallsten et al., 2005) assumes that people ignore the underlying stochastic process in the BART and simply choose a specific state to pump to (i.e., the target). Learning occurs by adjusting this target upward or downward, depending on the outcome of the previous trial. A successful Bayesian model (Wallsten et al., 2005) assumes that people have a prior belief about the expected payoff of different numbers of pumps and update this belief on the basis of the observed feedback. A main result from these cognitive modeling efforts is that people's performance in sequential decision problems appears to hinge on which assumptions they make about the underlying stochastic processes. To date, however, only stochastic processes leading to monotonic changes have been implemented. In the present article, we will thus investigate how and to what extent people learn in *nonmonotonic* decision problems.

Nonmonotonic Decision Problems

Nonmonotonic decision problems imply two key differences for learning relative to monotonic decision problems. First, repeated

increases and decreases in outcomes can result in a large number of different outcome distributions. For example, there could be a single maximum across the sequential states, or there could be multiple maxima with one being the global maximum and others being local maxima. Prior to learning, people are thus unaware of even the most basic properties of the underlying stochastic process in a nonmonotonic decision problem. In a monotonic decision problem, in contrast, it is evident a priori that there is only one maximum up to which the outcome will increase. Learning in nonmonotonic decision problems can therefore be expected to be challenging.

Yet, there is also a property conducive to learning. In particular, the repeated increases and decreases in outcomes lead to gradual changes in potential gains and losses, and people do not (necessarily) face the risk of a sudden and total loss but rather the risk of a gradual loss. The absence of the risk of a total loss may facilitate exploration. Assuming that people indeed explore extensively in nonmonotonic decision problems, one may expect them to make similar decisions in situations in which they receive partial feedback (i.e., thus facing an exploration–exploitation trade-off) and in situations with full feedback (thus rendering exploration unnecessary; see below).

The properties of nonmonotonic decision problems enable people to adopt one of several different learning strategies. Before explicating three of these learning strategies and the learning models representing them, we introduce a novel task called the Sequential Investment Task (SIT). We have designed this task such that it (a) can implement stochastic processes that lead to nonmonotonically changing outcomes and (b) permits the complexity of the underlying stochastic process as well as the amount of feedback provided to be varied.

The Sequential Investment Task

In the SIT, people sell shares in hypothetical stock markets. They initially "receive" 48 shares (each belonging to one of three stock markets; i.e., 16 shares per market), which they can sell over the course of virtual days. Each day represents one trial and offers the possibility to sell a share at one of 16 sequential selling points (see Figure 1). Thus, people need to decide whether to keep or sell a share at each specific time of a day, depending on the share's current price. If they do not sell a share over the course of a full day, the share will be automatically sold at the final selling point. The SIT can be parameterized on the following dimensions.

First, we can manipulate the *complexity of the stochastic processes*. In our experiments, we used Gaussian random walks to generate nonmonotonically changing outcomes. We varied complexity by implementing two types thereof: In "unimodal markets" (low complexity), the price of a share increases for some time before it decreases. Thus, on average, there is one selling point per market that yields the highest selling price—a market's "peak selling point." In "bimodal markets" (high complexity), the price of a share increases and decreases twice. Thus, two maxima exist per market, one being the global maximum that, on average, yields the highest selling price (i.e., the peak selling point) and one being a local maximum, which yields a lower selling price. In bimodal markets, it is thus arguably more difficult to discriminate the peak selling point from other selling points.

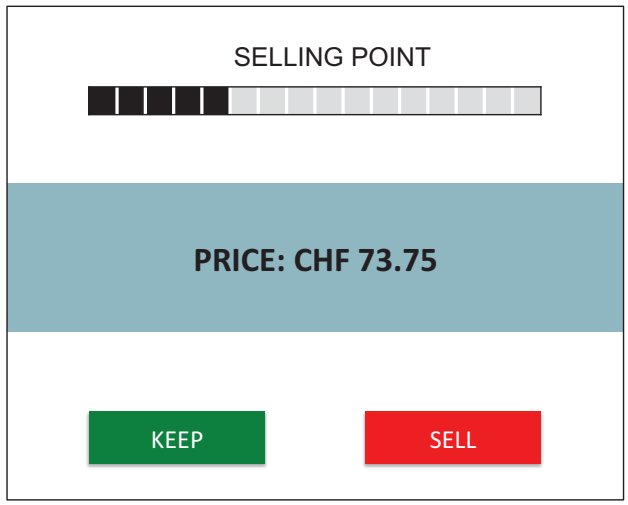

Figure 1. Screen displaying the Sequential Investment Task (SIT): At the top, a bar displayed progress across the 16 selling points in the virtual day. In the middle, participants saw the current price of the share, overlaid on a colored area that represented one of the three markets. At the bottom, the two choice options were displayed: to keep the share (left) or to sell the share (right). Participants indicated their choices by pressing a key marked with the respective color on the left or right of the keyboard. CHF = Swiss francs. See the online article for the color version of this figure.

Second, irrespective of the type of market (unimodal vs. bimodal), we can manipulate the *degree of exploration* that is required to observe the peak selling point. The three markets we implement in each of our experiments differ regarding whether the peak selling point is early, in the middle, or late in the selling period. The degree of exploration describes the number of selling points a person explores within one trial: For instance, only up to selling point 10 or all the way up to selling point 16. Note that another level of exploration in the SIT refers to how many of the 16 trials (i.e., the 16 shares per market) serve either as exploration trials (in which a person sells a share at a later selling point than he or she currently believes to be the subjectively best selling point) or as exploitation trials (in which a person sells the share at the subjectively best selling point).

Third, the SIT permits to provide either *partial* or *full feedback*. In the partial-feedback condition, people receive feedback about the prices of the shares only up to the point at which a share is sold. In the full-feedback condition, they also receive information about the development of a share's prices after it is sold. With partial feedback, people thus face an exploration–exploitation trade-off that is removed with full feedback.

Three Possible Learning Strategies

In nonmonotonic decision problems such as the SIT, there are at least three different learning strategies that people may apply. The first is to merely learn the state that yields the highest outcome, regardless of the outcomes of the other observed states. Learning occurs by updating the subjectively best state from trial to trial, depending on which of the observed states yielded the highest outcome in the previous trial. This strategy does not imply that

people assume the outcomes to be sequentially interdependent, and it does not build a full representation of the outcome distribution across the observed states. Therefore, it may be considered relatively parsimonious and cognitively not overly taxing. To represent this learning strategy, we use a target model as previously suggested for the BART (Wallsten et al., 2005); for similar models, see also Busemeyer and Myung (1992); Rieskamp, Busemeyer, and Laine (2003); and Selten and Stoecker (1986).

The second learning strategy is to learn a full representation of the outcome distribution across all observed states. Learning occurs by updating the subjective values (SVs) of the observed states on the basis of the feedback received. This strategy may be plausible in nonmonotonic decision problems for at least two reasons: On the one hand, it is not clear a priori that there is a single maximum, because the outcomes can increase and decrease repeatedly. Thus, people may aim to learn how attractive the different states are relative to each other, which requires a separate representation of all of them. On the other hand, the outcomes at each specific state in the SIT vary continuously because they are generated incrementally by a Gaussian random walk. Unlike in the BART, in which a specific state always results in exactly the same outcome (unless the balloon explodes at this particular state and the trial ends, resulting in a payoff of 0), in the SIT a specific state leads to a distribution of outcomes across trials. Therefore, people may represent each state as an independent random variable rather than as a state in a series of interdependent states. Like the first strategy, this strategy does not imply that people assume the outcomes to be sequentially interdependent, but it builds a full representation of the outcomes of all observed states. We implement a reinforcement-learning (RL) model to represent this second strategy, inspired by models that have received substantial support in the learning literature, such as in repeated-choice and *n*-armed bandit paradigms (Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Erev & Roth, 1998; Rieskamp, 2006; Yechiam & Busemeyer, 2005).

Finally, the third learning strategy is to evaluate the *actions* available at the different states of the sequential decision problem (i.e., whether to proceed to the next state or to stop at the current state). When evaluating a specific action, people may take into account not only the short-term consequences (i.e., the outcome that can be achieved in the next state) but also the long-term consequences (i.e., the outcomes that can be achieved in subsequent states). Such a learning strategy seems particularly plausible in nonmonotonic decision problems, where success may require learning that it is worth enduring a short-term decrease in outcomes and wait for outcomes to increase in subsequent states. Thus, this strategy not only builds a full representation of all observed states and the respective actions but also implies that people assume the outcomes to be sequentially interdependent. To represent the third learning strategy, we implement a variant of a Q-learning (QL) model (i.e., a temporal-difference learning model; Lohrenz, McCabe, Camerer, & Montague, 2007; Sutton & Barto, 1998), which has been proposed for sequential decision making, in particular.

The three models we implement are attractively simple but, at the same time, allow us to discriminate between learning strategies that are fundamentally different. Nevertheless, this set of models is not exhaustive and should be extended in future research (e.g., to include models that make detailed assumptions about human memory processes; see Gonzalez & Dutt, 2011).

Target model. According to the target model, people do not form SVs for all available selling points. Instead they focus on a single selling point (i.e., the target *T*) that they assume to yield the highest outcome. Learning takes place by using the feedback from the previous trial to update the target. In the SIT, the target is updated by shifting the current target in the direction of the selling point with the highest payoff in the previous trial.

We assume that people start with an initial target of selling the share in the middle of the selling period; that is, at selling point 8 (out of 16). Although the model assumes a single target, the choice of a selling point is probabilistic, so that all selling points may be chosen with a certain probability. The probability of choosing a specific selling point is a function of the distance of the selling point to the target. More specifically, we use a normal distribution with a mean of the target and a standard deviation as a free parameter to determine the choice probabilities. The probability of selling a share at selling point *s* (in market *M* and trial *t*), given the current target *T*, is defined as

$$p_{M,t}(s, T) = N(s + .5;\ T, \sigma) - N(s - .5;\ T, \sigma), \tag{1}$$

where $N()$ is the cumulative normal distribution function with a mean of T and a standard deviation of σ as a free parameter of the model specifying the sensitivity in the choice process.

After a person has experienced the optimal target of the current share (i.e., the selling point with the highest outcome), the target for the next trial $t + 1$ is updated by

$$T_{M,t+1} = T_{M,t} + \beta \cdot (T_{M,t,optimal} - T_{M,t}), \tag{2}$$

where $T_{M,t}$ is the current target, β a learning rate parameter and $T_{M,t,optimal}$ is the selling point with the highest payoff in trial *t*.

The following example illustrates the target model: In the first trial, a person decides to sell a share at selling point 8. It turns out that the highest outcome within selling points 1 to 8 would have been obtained at selling point 4. In this case, the person would shift the current target (selling point 8) toward selling point 4 and—assuming a learning rate of .5—set selling point 6 as the target for the next trial. In the next trial, the person would most likely sell the share at selling point 6. The target model assumes that people make a decision at the beginning of each trial, which is then carried out sequentially.

Reinforcement-learning model. The RL model assumes that people form SVs for each selling point in each market, which are updated on the basis of feedback. The probability of selling a share at a specific selling point is a probabilistic function of its SV relative to the SVs of the other selling points. The probability of selling a share at selling point *s* in market *M* and trial *t* is determined by an exponential choice rule defined as

$$p_{M,t}(s) = \frac{\exp(\theta_{RL} \cdot SV_{M,t}(s))}{\sum_{j=n}^{N=16} \exp(\theta_{RL} \cdot SV_{M,t}(j))}, \tag{3}$$

where θ_{RL} is a sensitivity parameter that determines how sensitively a person responds to differences in the SVs and N is the number of available choice options (i.e., the 16 selling points). After a decision, the feedback is used to update the SVs of all observed selling points as follows:

$$SV_{M,t+1}(s) = SV_{M,t}(s) + \delta_{RL} \cdot (r_{M,t}(s) - SV_{M,t}(s)), \quad (4)$$

where δ_{RL} is the learning rate parameter and $r_{M,t}(s)$ is the experienced or observed potential reward of selling the share at selling point s.

For instance, if a person decides to sell the first share at selling point 8 in the first trial (in which all SVs = 0), all the SVs of selling points 1 to 8 are updated with the observed potential selling prices. For example, if the price increased from CHF 72 (the initial price) to CHF 74.4 (the price at selling point 1; i.e., a gain of CHF 2.4), a person would update the SV for selling point 1 from 0 to 1.2 (assuming a learning rate of .5). Due to the updated SVs, different choice probabilities will result for the next trial. The RL model—like the target model—assumes that people select a selling point at the beginning of each trial and then sequentially implement the respective decision.

Q-learning model. This model assumes a sequential process in a narrower sense. It evaluates the choice options; that is, the action to "keep" or to "sell" a share at each state of the decision problem. The states are defined as the different selling points. Each state–action pair within one market is assigned a SV (i.e., a *Q-value*). At each state of the sequence, the model then chooses one action probabilistically as a function of the actions' Q-values.

A main characteristic of the QL model is that an action is evaluated not only in terms of its immediate consequences but also of its future consequences. Nevertheless, the immediate consequences are used directly after a decision to update the Q-values of the state–action pairs. In problems that permit backward and forward transitions (e.g., in networks), Q-learning benefits from this feature in terms of increased learning speed. We initialize the Q-values for all state–action pairs such that the initial probability of selling a share in a given market is equal for all 16 selling points.

At each selling point, the model applies the exponential choice rule described above (Equation 3, using the sensitivity parameter θ_{QL}) to decide whether to keep or sell a share at the current state, with the two available Q-values of the current state used as an input. After a decision, the model immediately updates the Q-values of the respective actions: If the decision was to continue (i.e., to keep the share), the Q-values for both continuing and stopping are updated (because, by then, the potential outcome of stopping is known). If the decision was to stop, only the Q-value for stopping is updated. The updating of the Q-values occurs according to

$$Q_{M,t+1}(s, a) = Q_{M,t}(s, a) + \delta_{QL} \cdot (r_{M,t}(s) + \gamma \cdot \max(Q_{M,t+1}) - Q_{M,t}(s, a)), \quad (5)$$

where $Q_{M,t}(s, a)$ are the SVs of the actions a (to keep or sell the share) in market M at selling point s, δ_{QL} is a learning parameter, r is the immediate reward from selling the share or from keeping it (in the later case, the immediate reward is 0), and γ is a prospective discounting parameter for $\max(Q_{M,t+1})$, which is the best possible Q-value available in the next state. In order to keep the number of free parameters over the models constant, we fix $\gamma = 1$.

To illustrate, let us assume that at selling point 1 the evaluation of the action "keep" is higher than that of the action "sell." Thus, the model may probabilistically predict that the person proceeds to selling point 2. The action of keeping the share at selling point 1 is then immediately updated by taking into account the new price of the share (i.e., the action is rewarded if the new price is higher but punished if the new price is lower). Similarly, the action of selling the share at selling point 1 is updated with the previously observed selling price. This process continues sequentially until the evaluation of a "sell" action is higher than the evaluation of a "keep" action at a specific selling point, and the share is sold.

For all selling points greater than 1, the QL model makes more choice predictions than the target model or the RL model. For example, to sell the share at selling point 10, the QL model first has to make nine "keep" decisions and then one "sell" decision. Models that assume such a step-by-step process generally have the following caveat: If the underlying stochastic process exhibits an early local maximum, the model may get stuck and stop. The reason is that the incremental rewards after a local maximum temporarily become smaller or even negative (until the next, global maximum approaches and the incremental rewards start to become positive again). In other words, step-by-step decision models may stop at a local maximum because continuing does not appear promising in the short term. The QL model aims to "look into the future," however, and takes into account the Q-values that will be available in the next state when making a decision. It thus reduces the probability of getting stuck at early local maxima.

Summary of model characteristics. The three learning models differ in various respects: First, the representation of the available options (i.e., the selling points in the SIT) differs substantially between the models. The target model represents only the subjectively best option, along with a measure of imprecision. It does not take different outcome distributions into account and could, for example, be misled in a bimodal market because it cannot directly compare the global and the local maximum. The RL model builds a representation of the full distribution of the options' outcomes, and every single selling point is updated independently. The RL model is thus able to represent different stochastic processes underlying the markets' outcomes and should be able to capture both the unimodal and bimodal markets' outcome distributions adequately. Reinforcement-learning models have been shown to successfully describe learning strategies in *n*-armed bandit problems (Ahn et al., 2008; Busemeyer & Stout, 2002; Erev & Roth, 1998; Yechiam & Busemeyer, 2005). Yet, it has not yet been shown whether people can also rely on such a learning strategy in the context of nonmonotonic decision problems. Finally, the QL model does not build representations of the options at all but rather representations of the actions available in each state (i.e., to keep or to sell a share at a specific selling point). Nevertheless, this mechanism can also lead to a representation of the full distribution of outcomes provided enough learning experience.

Second, the target model and the RL model update expectancies only after a final decision is made (i.e., when a share in the SIT is sold), whereas the QL model involves an immediate updating process that occurs after every single decision in the sequence. The more frequent updating process of the QL model leads to high performance in situations where each state offers multiple actions; that is, where several different transitions between states are possible (e.g., forward and backward transitions between different nodes of a network). In the case of the SIT, however, where only binary forward transitions are possible, the frequent updating process of the QL model could also imply a cost: To update the

expectancies of all state–action pairs, the QL model may require relatively extensive learning opportunities, relative to the RL and target models with their simpler updating processes.

Taking into account the different representations that the models can adopt, as well as the different updating frequencies they imply, we assume that the target model describes the learning strategy that is cognitively the least demanding, followed by the RL model and the QL model. Which learning strategy people eventually adopt, however, may depend not only on how cognitively demanding a strategy is but also on whether it successfully solves the decision problem. Specifically, in the environment of unimodal markets, all strategies may perform similarly, because there is only a global maximum. However, in the environment of a bimodal market, a strategy able to compare several maxima may perform better. In addition, how people learn may depend on the amount of feedback they receive about the environment. For example, full feedback may make the properties of an environment more evident and could thus increase the chance that people select a learning strategy that is better suited for an environment.

Experiment 1

In Experiment 1, we investigated three questions: First, to what extent can people improve their performance in nonmonotonic decision problems by learning? Specifically, do people learn to sell the shares at the selling points that on average yield the highest outcomes in the different markets of the SIT (i.e., at the peak selling points)? Second, does partial as compared to full feedback pose similar challenges to exploration as in monotonic decision problems (due to asymmetric error correction), or does the different risk profile (i.e., the risk of a gradual rather than a sudden loss) in nonmonotonic decision problems attenuate this gap? Third, which of the proposed computational models best describes the possible learning strategies in nonmonotonic decision problems, and does the selection of learning strategies depend on the type of feedback?

To keep complexity relatively low in Experiment 1, we implemented stochastic processes that generate a single maximum in each of three markets ("unimodal markets"). The peak selling points in the three markets were 4, 11, and 14, respectively. Thus, the three markets required increasing degrees of exploration within the selling period to find the peak selling points. Half of the participants were provided with partial feedback and thus faced an exploration–exploitation trade-off; the other half were provided with full feedback and thus faced no such trade-off.

Method

We recruited 63 participants from a subject pool at the University of Basel and randomly assigned them to one of two between-subjects conditions: partial and full feedback. Participants' age ranged from 18 to 54 years, with a mean of 24.4 years ($SD = 7.2$ years); 58 (92%) of them were female, and 5 (8%) were male. They received a voucher of CHF 15 (CHF 1 $\approx$ US$1.14) for participating plus a performance-contingent bonus: Starting with an initial credit of CHF 3, participants were paid according to their mean performance. For every percentage point increase or decrease in the share price, they earned or lost CHF 0.2 (e.g., if the initial price of a share was CHF 100 and a participant sold the share for CHF 110, the increase of 10% led to a bonus of CHF 2).

Participants were instructed that they were going to sell shares in three different, color-coded stock markets. They were informed that each market had a specific selling point that, on average, yielded the highest selling price. We explained that the prices of all shares in each market followed a similar price pattern due to an external event. Participants were further informed that they would receive 16 shares for each of the three markets (i.e., a total of 48 shares) and that they could sell each of these shares during a virtual day. Each day consisted of 16 selling points that required a binary decision on whether to keep or to sell the current share. If not sold earlier, a share was automatically sold at selling point 16.

After a practice trial, participants were presented with the first of the 48 shares. They were asked to sequentially decide whether to keep or to sell the share for the current price at the 16 different selling points in the hypothetical day (see Figure 1) and had 2 seconds to make each decision (by pressing two different keys on the keyboard). Informed by a short pilot study, we set this time limit such that participants were neither under particular time pressure nor bored. At 1353 ms, the average response time proved to be substantially below this limit. After each decision, the price remained visible until a total of 4 seconds had elapsed so that participants had enough time to register the current price. If participants did not press a key for 2 seconds, this was interpreted as a "keep decision" and the program moved on to the next selling point. When participants in the partial-feedback condition sold a share, the trial was over and a new share was randomly drawn from one of the three markets. When participants in the full-feedback condition sold a share, the prices of the remaining selling points were sequentially displayed (each price for 2 seconds), without the participant taking any action, before a new share was randomly drawn.

The initial price of each share was defined as a random draw from a Gaussian distribution with $M = 70$ and $SD = 5$, and a Gaussian random walk determined the increments of the share's price per selling point. Specifically, the increments were random draws from a Gaussian distribution with $M = 2$ and $SD = 2.3$. After the peak selling point of the market had been reached, these increments were multiplied by −1 to produce decreasing prices. Figure 2 shows the resulting prices of the 16 shares in each market.

Results

Learning curves and selling points. Figure 3 shows participants' learning curves over the course of the trials (i.e., the 16 shares participants could sell in each market), separately for the three markets and the two feedback conditions. On average, participants sold the first shares per market approximately in the middle of the selling period (i.e., around selling point 8). Over the course of trials, however, they learned to sell closer to the peak selling points of the markets. In market 1 with a peak selling point of 4, participants in both conditions (partial and full feedback) approximated this point closely. In markets 2 and 3 with peak selling points of 11 and 14, respectively, only participants in the full-feedback condition—and not those in the partial-feedback condition—approximated the peak selling point across trials.

To compare the learning curves on a quantitative level between conditions and markets, we determined the absolute deviation of participants' selling points from the peak selling points. To distinguish between levels of learning opportunity, we created four

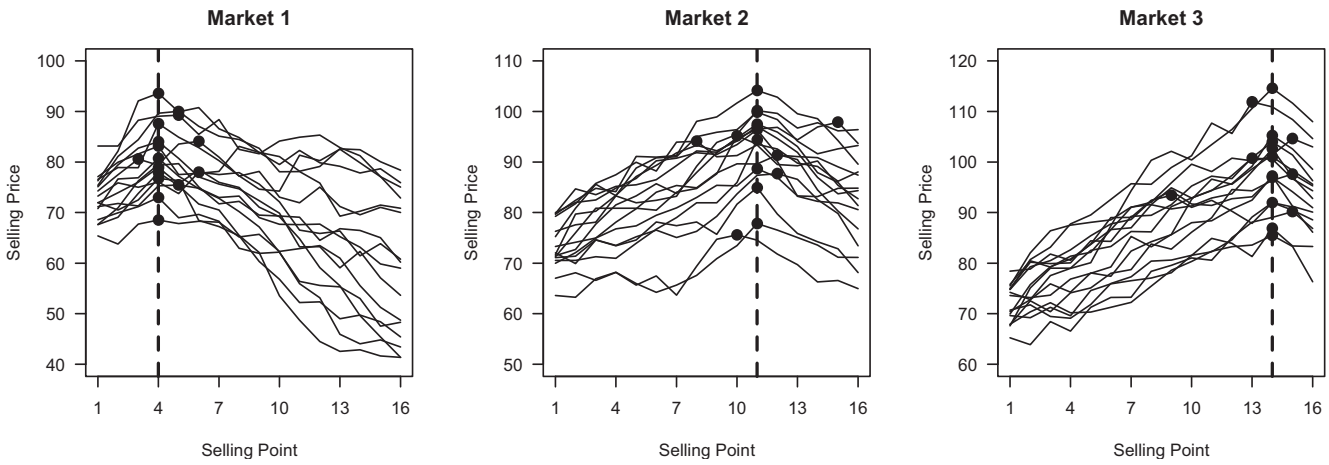

Figure 2. The prices of the 16 shares in the three markets of Experiment 1, as a function of the 16 selling points. The dashed vertical lines depict the peak selling point in each market (defined as the selling point that on average yields the highest outcome). Dots represent the highest selling price for each of the shares.

blocks of four consecutive shares. Figure 4 (top panel) shows the mean absolute deviations separately for the two conditions and the three markets over the four blocks of shares. In a three-way analysis of variance with two within-subject factors (market [1–3] × block [1–4]) and one between-subjects factor (condition [partial vs. full feedback]), there was a main effect for block, with $F(1, 744) = 14.5$, $p < .001$. That is, over the course of trials, participants in both conditions learned substantially. As can be seen in Figure 4, overall the absolute deviations from the peak selling points became increasingly smaller across blocks. We also found a significant main effect for condition, $F(1, 744) = 46.3$, $p < .001$, with smaller deviations in the full-feedback condition. Finally, there was a significant interaction between condition and market, $F(1, 744) = 44.2$, $p < .001$, implying that the advantage of participants in the full-feedback condition over those in the partial-feedback condition was distinct to specific markets. For example, in market 1, where the peak selling point was early, deviations in both conditions were comparably small. However, in markets 2 and 3, where the peak selling points were late, deviations were smaller in the full-feedback condition (see Figure 4).

In sum, it made a substantial difference whether participants experienced partial or full feedback in the markets with late peak selling points, in which more exploration was needed to observe them. Next, we investigated the learning processes in more detail, in particular which of the proposed models captured them best.

Model comparison. The free parameters of the three models were separately estimated for each participant using maximum likelihood techniques. The models were used to predict participants' choices in the next trial, and the feedback they received was used to update the expectancies of the models. Thus, we used a one-step-ahead estimation method. Figure 3 shows the average learning curves that resulted from the medians of the estimated model parameters. The target model performed well in the full-feedback condition but did not describe the learning processes accurately in the partial-feedback condition. The RL model captured the empirical learning processes reasonably well in both the partial and the full-feedback conditions. The QL model was, on average, relatively poor in describing the learning processes, particularly in markets 2 and 3 with partial feedback. Participants' learning processes were quicker than the QL model was able to describe them, given the available learning opportunity. In addition, the QL model learned that it is not worth waiting until the end of the sequence.

Next, we evaluated the models against a baseline model that predicts each selling point with equal probability using the Bayesian information criterion (BIC; Schwarz, 1978). The BIC takes a model's goodness-of-fit and complexity in terms of adjustable parameters into account. The learning models can do better than the baseline model only if they are able to make sensible use of the feedback in describing the observed learning processes. As can be seen from the distributions of model fits across participants (see Figure 5), all models performed better than the baseline model, especially in the full-feedback condition. The estimated model parameters are displayed in Table 1. Due to the different nature of the models, these parameters cannot be compared directly across the models; however, they can be compared between conditions within each model. The learning rates can be interpreted as the extent to which people adjust their target (target model), their SVs of the different selling points (RL model), or their SVs of the available actions at each state (QL model) from trial to trial. High learning rates imply a high degree of adjustment, which leads to fast learning at the beginning of the learning process but could result in "overadjustment" in the long run (i.e., the learning process could switch too strongly between different options). Low learning rates, in contrast, imply slower but more fine-grained adjustment and learning. Across all three models, learning rates were higher in the partial-feedback condition than in the full-feedback condition. In other words, the adjustment process in the full-feedback condition was more fine-grained, whereas in the partial-feedback condition it was more sensitive to the most recent trial (and may thus have resulted in overadjustment).

The choice sensitivity parameter implied higher sensitivity in the full-feedback condition than in the partial-feedback condition for all models (note that the target model's sensitivity parameter has to be interpreted inversely; see Table 1). High choice sensitivity means that participants selected the subjectively best option

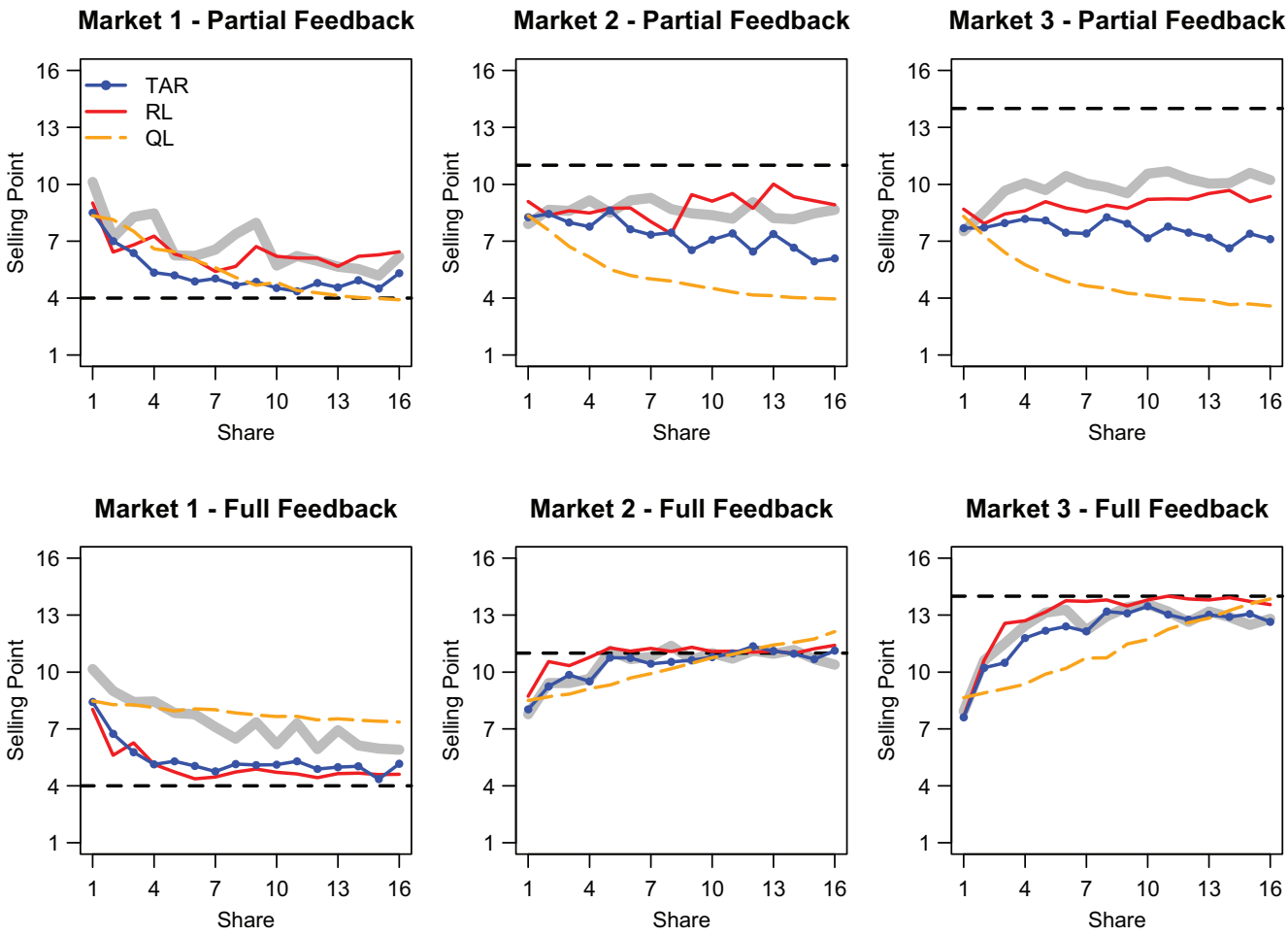

Figure 3. Learning curves in Experiment 1: The thick lines represent participants' average selling points, the dashed horizontal lines depict the peak selling points in each market, and the remaining three thin lines represent the models' learning curves resulting from using the medians of the estimated parameters. The top row displays participants' learning curves in the partial-feedback condition; the bottom row displays those in the full-feedback condition. TAR = target model; RL = reinforcement-learning model; QL = Q-learning model. See the online article for the color version of this figure.

with high probability. Thus, participants in the full-feedback condition made more deterministic decisions given their learning experience.

Participant classification. Next, we examined which learning model performed best in describing the learning process of each participant, separately for the two feedback conditions. To this end, we used Bayesian model weights based on the models' BICs (Lewandowsky & Farrell, 2010; Raftery, 1995), which indicate the strength of evidence for the best model relative to the other models. When none of the three learning models were better than the baseline model, the respective participants remained unclassified. As Figure 6 shows, in the condition with partial feedback, each of the three models best described the learning processes of a comparable proportion of participants: 41% were best described by the target model, 31% by the QL model, and 25% by the RL model. A small proportion of participants (3%) remained unclassified. In the condition with full feedback, the learning processes of the majority of participants (i.e., 68%) were assigned to the RL model, followed by the target model with 16%, and the QL model with 6%. A proportion of 10% of participants remained unclassified. In sum, although no model clearly performed best in the partial-feedback condition, the RL model best described the learning processes of most participants in the full-feedback condition.

Discussion

Despite several challenges to learning (e.g., the different markets were presented in alternating order; see Lejuez et al., 2002), we found that participants learned to adjust their selling points toward the peak selling points over the course of the trials. The feedback conditions had a substantial influence on participants' learning processes, however. That is, when the peak selling points were late in the selling period, participants in the partial-feedback condition did not manage to closely approximate the peak selling point. One reason is that they did not explore the full range of the selling period and therefore never experienced the peak selling points in these markets. It was only in the

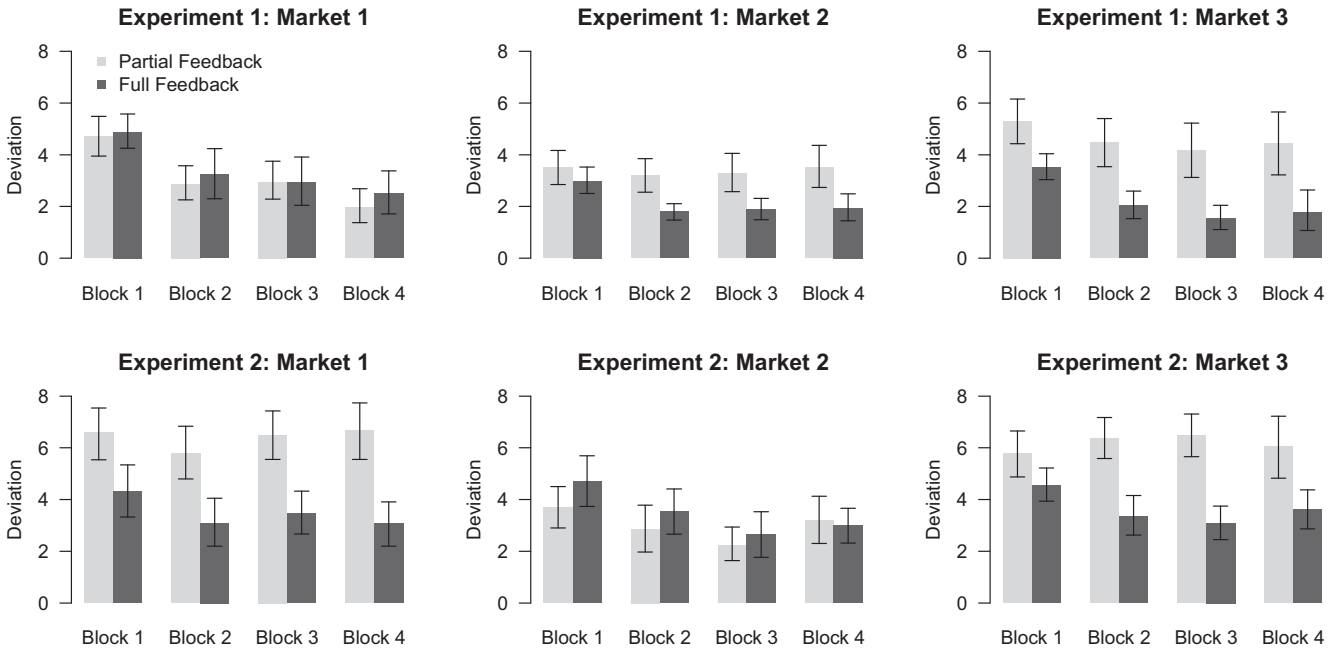

Figure 4. Mean deviations of participants' selling points from the peak selling points, separately for the two conditions and the three markets. The top row shows the results of Experiment 1; the bottom row shows the results of Experiment 2. Each block represent four consecutive shares; namely, shares 1–4 (Block 1), shares 5–8 (Block 2), shares 9–12 (Block 3), and shares 13–16 (Block 4). Error bars show 95% confidence intervals.

full-feedback condition that participants adjusted their selling points toward the peak selling points in all three markets. Given that exploration in a nonmonotonic decision problem such as the SIT is not overly risky, it is surprising that the feedback condition affected participants' choices so strongly.

Our modeling analysis suggests that learning was qualitatively and quantitatively influenced by the two feedback conditions. On a qualitative level, each model best described the learning processes of about a third of the participants in the partial-feedback condition. In the full-feedback condition, in contrast, the RL model best described the learning processes of most of the participants. The good performance of the RL model in the full-feedback condition suggests that participants absorbed the additional information that was presented. Specifically, a crucial feature of this model is that it leads to a complete representation of all selling points. The QL model could also have led to a full representation, but it would apparently have required more learning opportunities, given the relatively complex updating process. A potential reason why the RL model's performance was poorer in the partial relative to the full-feedback condition is that the unobserved states cannot be updated. Therefore, the model tends to predict earlier selling points with much higher likelihood, whereas the target model predicts later selling points with some likelihood even in the case of partial feedback. On a quantitative level in terms of the estimated parameters, participants in the full-feedback condition adopted a more fine-grained learning process (smaller learning rates) and, at the same time, were more deterministic when making their decisions (higher choice sensitivity).

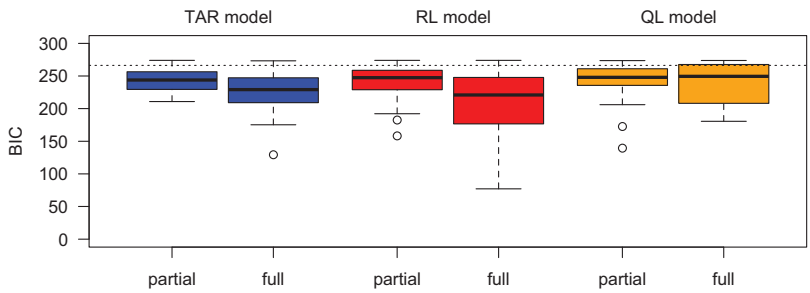

Figure 5. Distribution of model fits in Experiment 1: Box plots depict the distributions of fits per model and feedback condition, measured with the Bayesian information criterion (BIC). The dashed horizontal line depicts the BIC of a baseline model choosing a selling point with equal probability. TAR = target; RL = reinforcement learning; QL = Q-learning. See the online article for the color version of this figure.

Table 1
Parameter Estimates of the Learning Models (Medians)

Model	Parameter	Experiment 1 Partial feedback	Experiment 1 Full feedback	Experiment 2 Partial feedback	Experiment 2 Full feedback
Target model					
Learning rate	β	.39	.33	.50	.45
Sensitivity[a]	σ	3.6	2.7	4.5	5.6
RL model					
Learning rate	δ_{RL}	.34	.18	.37	.18
Sensitivity	θ_{RL}	.20	.50	.12	.19
QL model					
Learning rate	δ_{QL}	.09	.02	.45	.03
Sensitivity	θ_{QL}	.34	2.76	.33	.25

Note. RL = reinforcement learning; QL = Q-learning.
[a] The target model's sensitivity parameter has to be interpreted inversely, as σ stands for the imprecision of the choice. Thus, lower values represent higher sensitivity.

Experiment 2

In the unimodal markets with a single maximum in Experiment 1, the same performance could theoretically be achieved either by learning a full representation of all selling points (as described by the RL or QL model) or by learning only the subjectively best selling point (as described by the target model). In Experiment 2, we increased the complexity of the stochastic processes and implemented bimodal markets with two maxima, one being a global maximum (the peak selling point) and one being a local maximum.

In a bimodal market, the target learning strategy might be expected to perform worse because it could be misled by the local maximum. In contrast, the RL model and the QL model can solve learning problems with several maxima as they result in a complete representation of all states. Thus, if people select learning strategies by the anticipated performance, they may rely more on the RL or QL model in situations with several maxima, which they may detect at an early learning state. This prediction should in particular hold in learning situations with full feedback, in which it is easy to detect that several maxima exist. In Experiment 2, we will test this prediction and the extent to which people adapt their decision strategies to the type of environment (Rieskamp, 2006; Rieskamp & Otto, 2006).

Method

We recruited 63 participants from a subject pool at the University of Basel and randomly assigned them to one of two between-subjects conditions: partial and full feedback. Their age ranged from 17 to 49 years, with a mean of 24.34 years (SD = 6.5 years); 53 (84%) of them were female, and 10 were male. The procedure of Experiment 2 was identical to that of Experiment 1. The only difference was that the stochastic processes used in the three markets resulted in a global as well as a local maximum. The initial price of each share in Experiment 2 was defined as a random draw from a Gaussian distribution with M = 75 and SD = 5, and a Gaussian random walk determined the increments of the share's price per selling point. The increments were random draws from a Gaussian distribution such that the global maxima for the three markets were at selling points 13, 3, and 15, whereas the local maxima were at the selling points 4, 14, and 5. Figure 7 shows the resulting prices of the 16 shares in each market.

Results

Learning curves and selling points. Figure 8 shows participants' learning curves over the course of the trials (i.e., the 16 shares participants could sell in each market), separately for the three markets and the two feedback conditions. On average, participants again sold the first shares per market approximately in the middle of the selling period (i.e., around selling point 8). In the

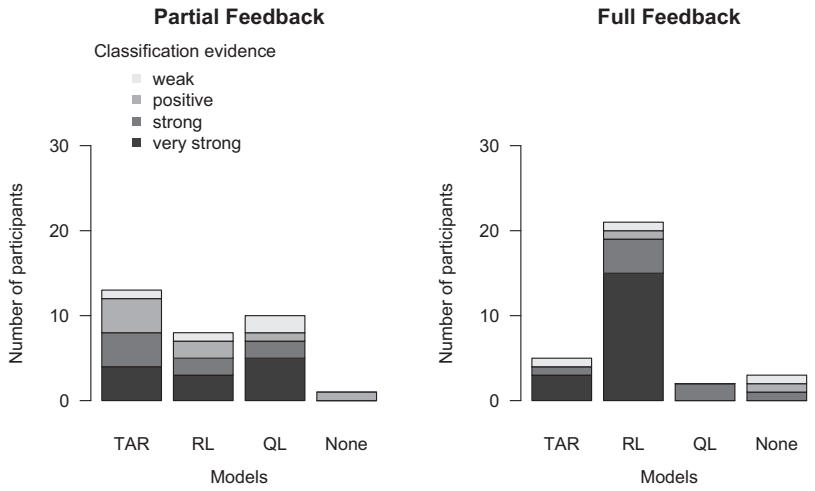

Figure 6. Classification of participants' learning processes in Experiment 1 according to Bayesian model weights (Lewandowsky & Farrell, 2010; Raftery, 1995). The left (right) chart shows the number of participants in the partial (full) feedback condition whose learning processes were best described by each of the three models, including the relative evidence for the best fitting model over the other models. TAR = target model; RL = reinforcement-learning model; QL = Q-learning model.

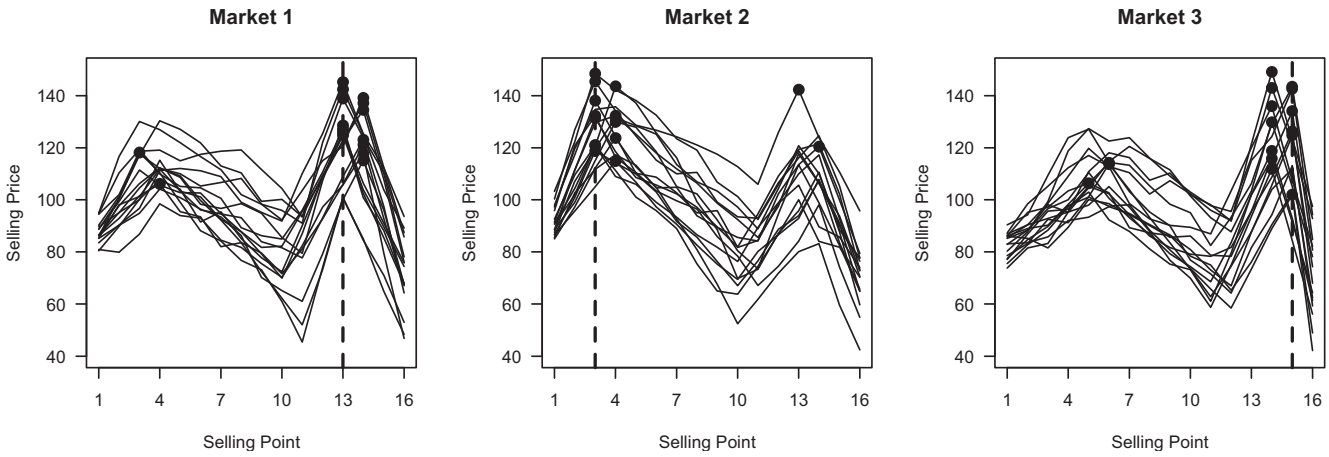

Figure 7. The prices of the 16 shares in the three markets of Experiment 2, as a function of the 16 selling points. The dashed vertical lines depict the peak selling point in each market (defined as the selling point that on average yields the highest outcome). Dots represent the highest selling price for each of the shares.

partial-feedback condition, participants shifted their selling points to the peak selling point only in market 2, where the global maximum was early in the selling period. In the other two markets, they did not shift their selling points toward the peak selling points. In the full-feedback condition, in contrast, participants shifted their selling points toward the peak selling point in all markets. Yet, compared with the corresponding condition in Experiment 1 (e.g., market 3, full-feedback condition, where the peak selling point was also late in the selling period), the learning process seemed slower in Experiment 2. The local maximum evidently influenced whether and how closely participants approximated the peak selling points.

We next determined the absolute deviation of participants' selling points from the peak selling points. As we did in in Experiment 1, we created four blocks of shares to distinguish between levels of learning opportunity. Figure 4 (bottom panel) shows the mean absolute deviations, separately for the two conditions and the three markets over the four blocks of shares. In a three-way analysis of variance with two within-subject factors (market [1–3] $\times$ block [1–4]) and one between-subjects factor (condition [partial vs. full feedback]), there was a main effect for condition, $F(1, 744) = 63.7$, $p < .001$, with smaller deviations from the peak selling points in the full-feedback condition. There was also a significant interaction between condition and block, $F(1, 744) = 4.9$, $p = .03$, implying that the learning differences between participants in the partial versus full-feedback condition were not the same over the four blocks of trials. This reflects the fact that participants in the full-feedback condition mainly learned from the first to the second block; that is, their deviations shrank early but did not change any further. The participants in the partial-feedback condition, however, barely learned anything over the full course of blocks (see Figure 4).

Model comparison. As we did in Experiment 1, we estimated the free parameters of the three models using maximum likelihood techniques, separately for each participant. Figure 8 shows the average learning curves that resulted from the medians of the estimated model parameters. The RL model captured the average learning curves reasonably well in almost all markets and conditions. In market 1 of the partial-feedback condition, it described a slightly more pronounced shift to the local maximum as compared to the actual data. Overall, the target model described a very similar pattern; however, it captured the observed learning processes less precisely. The QL model had a tendency to shift toward the beginning of the selling period in all markets in the partial-feedback condition, and in the full-feedback condition, learning was slow and did not succeed in capturing the observed learning processes well. As Figure 9 shows, the RL model achieved the best fit (in terms of BICs), in both the partial and the full-feedback conditions. The target model and the QL model were less accurate but still better than the baseline model. Thus, at least on the aggregate level, the distribution of model fits suggests that the RL model outperformed the other two models in the more complex learning environment of Experiment 2.

The estimated model parameters are displayed in Table 1. As in Experiment 1, learning rates were higher in the partial-feedback condition, suggesting stronger (and possibly over-) adjustment during the learning process, whereas the lower learning rates in the full-feedback condition suggest a more fine-grained adjustment process. Overall, the learning rates matched the estimated parameters of Experiment 1 relatively closely, in particular for the RL model. However, presumably because of the bimodal outcome distributions, choice sensitivity in all models decreased markedly, describing more random choices (note that the target model's sensitivity parameter has to be interpreted inversely; see Table 1). That is, participants in Experiment 2 were not able to exploit their acquired knowledge as well as participants in Experiment 1 did.

Participant classification. We next examined which learning model was best able to describe the learning process of each participant, separately for the two conditions and using Bayesian model weights as in Experiment 1. In both the partial and the full-feedback conditions, the RL model best described the learning processes of the vast majority of participants (see Figure 10). In the partial-feedback condition, the RL model best described the learning processes of 85% of participants, followed by the QL model (9%) and the target model (3%). Another 3% of participants remained unclassified. In the full-feedback condition, the learning

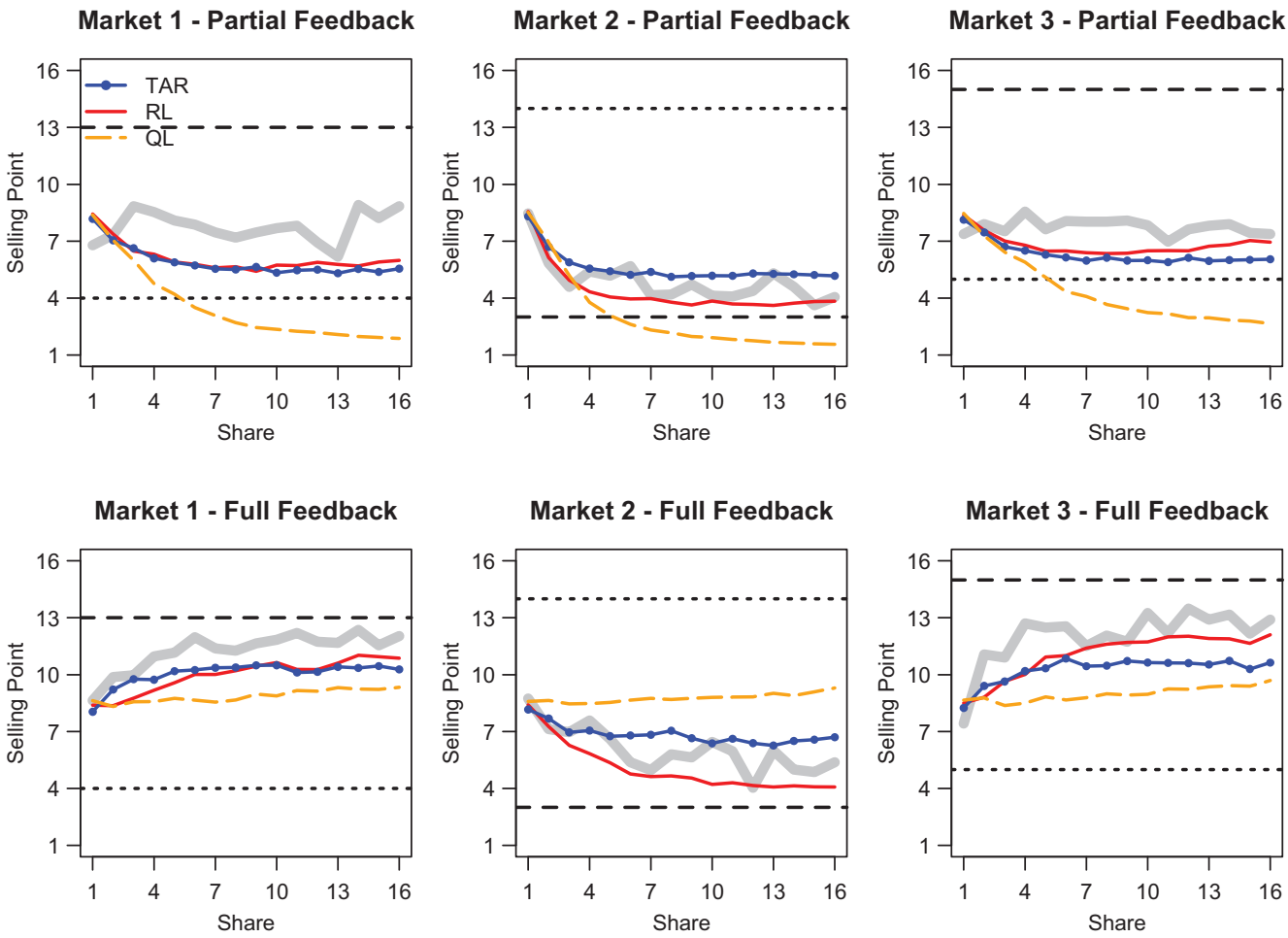

Figure 8. Learning curves in Experiment 2: The thick lines represent participants' average selling points, the dashed horizontal lines depict the peak selling points in each market, the dotted horizontal lines depict the local maxima in each market, and the remaining three thin lines represent the models' learning curves resulting from using the medians of the estimated parameters. The top row displays participants' learning curves in the partial-feedback condition; the bottom row displays those in the full-feedback condition. TAR = target model; RL = reinforcement-learning model; QL = Q-learning model. See the online article for the color version of this figure.

processes of all participants were best described by the RL model. The evidence for this classification in terms of Bayesian model weights was overall very strong (Raftery, 1995; Lewandowsky & Farrell, 2010). As predicted, the bimodal markets led participants to rely more frequently on a learning strategy that builds a full representation of all observed states.

Discussion

Experiment 2 introduced a more complex environment with a global and a local maximum. Participants in the partial-feedback condition did not adjust their selling points toward the peak selling points in markets where these occurred late in the selling period. Instead, they seemed to get stuck between the global maximum (at the end of the selling period) and the local maximum (at the beginning of the selling period). In the full-feedback condition, in contrast, participants approximated the peak selling points in all three markets.

The main goal in Experiment 2 was to investigate whether more complex stochastic processes lead people to adopt learning strategies that result in a full representation of all selling points. Although on the aggregate level, the differences between the RL model and the target model appeared relatively small, the classification analysis clearly indicated that participants' learning processes were best described by the RL model, in both the partial and the full-feedback condition. In addition, the proportion of participants best described by the target model decreased markedly from Experiment 1 to Experiment 2 (see partial-feedback condition, Experiment 1). Thus, participants in both conditions seem to have realized that the underlying stochastic processes can be better grasped with a learning strategy that represents all selling points. In the full-feedback condition, in which it was arguably easier to learn about the environment and which strategy was appropriate, the evidence for the RL model was even stronger than in the partial-feedback condition.

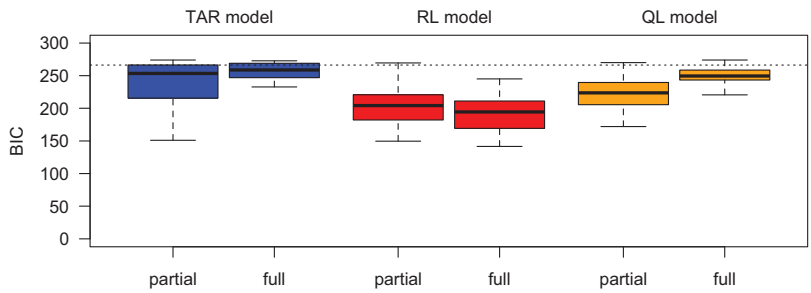

Figure 9. Distribution of model fits in Experiment 2: Box plots depict the distribution of fits per model and feedback condition, measured with the Bayesian information criterion (BIC). The dashed horizontal line depicts the BIC of a baseline model choosing a selling point with equal probability. TAR = target model; RL = reinforcement-learning model; QL = Q-learning model. See the online article for the color version of this figure.

General Discussion

In two experiments, we examined people's learning processes in nonmonotonic decision problems. Learning is challenging in this environment because no contextual information is provided about how the outcomes change across the sequence of decision options. We investigated how the complexity of the environment and different types of feedback affected people's decisions and their learning processes. We found that people chose selling points much closer to the markets' peak selling points with full feedback than with partial feedback. With partial feedback, participants faced an exploration–exploitation trade-off and their performance suffered, especially in an environment with high complexity. The cognitive modeling analysis showed that, in the majority of cases, a reinforcement-learning model best described participants' learning processes. This model assumes that people build a full representation of all options (selling points). Only in environments with low complexity and with partial feedback did a substantial proportion of participants seem to follow a simple learning process that focuses on acquiring the subjectively single best option. These findings suggest two key conclusions.

The Role of Feedback in Nonmonotonic Decision Problems

The extent to which people approximate the "sweet spot" in a nonmonotonic decision problem appears to depend largely on an interaction between the degree of exploration required to find this spot (e.g., a peak selling point later in the selling period requires more exploration in the SIT) and the amount of information provided via feedback (i.e., partial vs. full feedback). In particular, with a late peak selling point and only partial feedback, participants did not manage to postpone their selling decision long enough in the SIT. This effect was even more pronounced when outcome distributions were bimodal and involved (early) local maxima (Experiment 2). When participants experienced full feedback, however, they took advantage of the available information

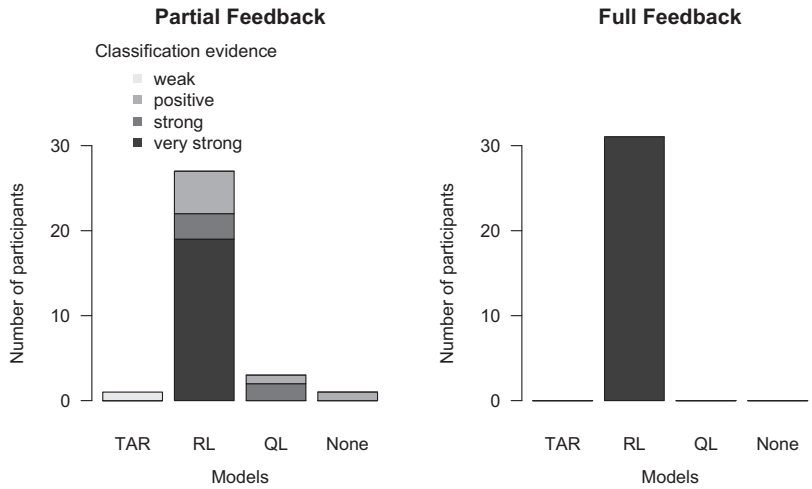

Figure 10. Classification of participants' learning processes in Experiment 2 according to Bayesian model weights (Lewandowsky & Farrell, 2010; Raftery, 1995). The left (right) chart shows the number of participants in the partial (full) feedback condition whose learning process were best described by each of the three models, including the relative evidence for the best fitting model over the other models. TAR = target model; RL = reinforcement-learning model; QL = Q-learning model.

and, with increasing learning experience, were able to postpone their selling decision.

Full information does not always translate into better performance, however: If people experience full feedback about options that are better most of the time but worse on average (i.e., skewed distributions), they will be attracted to suboptimal choice options (Yechiam & Busemeyer, 2006). Also, research on a related task has shown that people can make successful inferences from selective feedback using constructivist coding, equaling the performance of others receiving full feedback (Elwin, Juslin, Olsson, & Enkvist, 2007). However, in nonmonotonic decision problems as implemented in the SIT, hypothetical outcomes (i.e., forgone payoffs) clearly helped people to adjust their selling points toward the peak selling points. Similar observations have been made in other domains (Camerer & Ho, 2003; Chiu et al., 2008; Grosskopf, Erev, & Yechiam, 2006; Lohrenz et al., 2007).

In a nonmonotonic decision problem, there is not necessarily the risk of a sudden loss during exploration (unlike, for instance, in the BART); therefore, it is not trivial that a clear gap persisted between the choices in the partial- and full-feedback conditions. A possible reason for this gap is that risk aversion may play a role not only in paradigms such as the BART but also to some degree in the SIT. That is, participants may have been satisfied with a small initial increase in the selling price and did not want to run the risk of gradually losing what they had just gained. They therefore stopped exploration after finding a sufficiently attractive price (Einav, 2005). A lack of desire to explore has also been shown in other experienced-based decision tasks (Hau, Pleskac, & Hertwig, 2010; Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Pleskac, 2010).

What assumptions people in the partial-feedback condition may have made about the unobserved states remains an open question. Elwin et al. (2007) suggested that people deal with unobserved states by either using "positivist coding" or "constructivist coding." With positivist coding, people have no assumptions about the unobserved states, and decisions are exclusively based on the observed states. With constructivist coding, in contrast, people form some assumptions and expectations about the remaining states based on a particular rule. For the SIT, it is currently unclear what form such a rule would take. In future studies, one way to explore such rules would be to ask people to predict the prices that subsequent states will yield. On the basis of these predictions, it would be possible to infer people's expectations and some of the rules they may use for constructivist coding in nonmonotonic decision problems.

In sum, although the risk profile in the nonmonotonic decision problem of the SIT should render exploration less risky, people who faced an exploration–exploitation trade-off seemed to get stuck at early states of the decision problem during exploration—particularly if outcomes showed an early (momentary) downtrend.

Learning Strategies in Nonmonotonic Decision Problems

The results of the model comparison suggest that people tend to rely on a reinforcement-learning strategy in nonmonotonic decision problems; that is, that they learn a representation of all observed states. In the SIT, it was only in the partial-feedback condition of Experiment 1, in which complexity was relatively low, that other learning strategies best described the learning processes of a sizable proportion of participants. Unlike the target model, the QL model is also capable of building a full representation of all decision options. However, its learning and updating processes were too slow to adequately describe participants' learning curves. The QL model might perform well with more learning opportunities or in sequential decision problems that permit more than just two actions at each state (i.e., to sell or to keep the share).

Further, the modeling results suggest that the estimated parameters may capture two different aspects: On the one hand, the feedback conditions had a substantial quantitative influence on people's learning processes. We compared the fitted parameters across the two conditions in each of the two experiments and found systematic differences regarding learning speed and choice sensitivity. On the other hand, there was variability within the separate conditions, reflecting interindividual differences. Future research may benefit from relating measures of cognitive abilities, such as working memory span (Lewandowsky, Oberauer, Yang, & Ecker, 2010) or processing speed (Wechsler, 1981), to the estimated model parameters and to the specific model by which a person is best described. In addition, it seems promising to adapt models that make additional assumptions about human memory processes in learning, such as instance-based learning models (e.g., Gonzalez & Dutt, 2011).

Limitations and Future Research

In some of the conditions (see Figures 3 and 8), the amount of learning observed was modest. Of course, learning in nonmonotonic decision problems is generally more demanding than it is in monotonic decision problems. But two other design aspects may also have hampered learning. First, in line with the procedure used in the BART (Lejuez et al., 2002), we presented shares from three different markets in a random, alternating order. This approach may have made learning even more challenging. Second, the drift-to-noise ratio in the Gaussian random walk, which determines the sequential changes in a share's price, may also directly influence learning. We aimed to set this ratio so that participants would have a reasonable chance to learn, without making learning too simple or too difficult. Other ratios may be examined in future studies.

The SIT can be easily parameterized, and future research could also adapt it to represent a true dynamic decision problem such that the prices of a share change dynamically as a function of the participant's (or even other participants') previous decisions across trials. This would allow researchers to investigate, for instance, the influence of competition as it occurs in the stock market, where individuals learn and make decisions in parallel.

Conclusions

Many real-world nonmonotonic decision problems (e.g., the stock market) do not rest on stable stochastic processes that people can easily learn about. Nevertheless, people may often aim to do so, and our stylized task illustrated some of the circumstances under which learning succeeds or is impeded. It has been argued

that people are able to adapt their decision strategies to the type of the environment (Pleskac, 2008; Rieskamp, 2006; Rieskamp & Otto, 2006). Our results suggest that people also adopt appropriate learning strategies in the context of nonmonotonic decision problems, depending on the complexity of the environment.

Yet, a challenging property of many sequential decision problems is that they require a trade-off between exploration and exploitation. Exploration in nonmonotonic decision problems such as the SIT does not carry the risk of a sudden and total loss. Nevertheless, our participants did not engage in extensive exploration; consequently, they did not find out about peak selling points occurring late in the sequence. This observation may partly explain why people tend to "buy high and sell low" (Benartzi & Thaler, 2007) in real stock markets: As positions typically disappear from a portfolio when they are sold, investors—unless they actively seek this information—lack feedback on prices that could have been obtained at later selling points. This asymmetric feedback may also prompt an asymmetry in error correction, reinforcing the tendency to sell earlier than later. If positions remained visible in the portfolio after a selling decision, investors could more easily learn about foregone gains and losses—and about potential trend reversals. Such modifications to the architecture of investment decisions may help some investors to hone their market timing skills by making them aware of the possibly undue influence of recent market trends.

References

Ahn, W. Y., Busemeyer, J., Wagenmakers, E. J., & Stout, J. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32,* 1376–1402. http://dx.doi.org/10.1080/03640210802352992

Benartzi, S., & Thaler, R. H. (2007). Heuristics and biases in retirement savings behavior. *Journal of Economic Perspectives, 21,* 81–104. http://dx.doi.org/10.1257/jep.21.3.81

Bouman, S., & Jacobsen, B. (2002). The Halloween indicator, "Sell in May and go away": Another puzzle. *American Economic Review, 92,* 1618–1635. http://dx.doi.org/10.1257/000282802762024683

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121,* 177–194. http://dx.doi.org/10.1037/0096-3445.121.2.177

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14,* 253–262. http://dx.doi.org/10.1037/1040-3590.14.3.253

Caldwell, K. (2013, August 27). "Sell in May" costs investors £50,000 in a decade. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/finance/personalfinance/investing/10268436/Sell-in-May-costs-investors-50000-in-a-decade.html

Camerer, C., & Ho, T. H. (2003). Experience-weighted attraction learning in normal form games. *Econometrica, 67,* 827–874. http://dx.doi.org/10.1111/1468-0262.00054

Chiu, P. H., Lohrenz, T. M., & Montague, P. R. (2008). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. *Nature Neuroscience, 11,* 514–520. http://dx.doi.org/10.1038/nn2067

De Bondt, W. P. (1993). Betting on trends: Intuitive forecasts of financial risk and return. *International Journal of Forecasting, 9,* 355–371. http://dx.doi.org/10.1016/0169-2070(93)90030-Q

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review, 114,* 177–187. http://dx.doi.org/10.1037/0033-295X.114.1.177

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science, 12,* 523–538. http://dx.doi.org/10.1287/orsc.12.5.523.10092

Einav, L. (2005). Informational asymmetries and observational learning in search. *Journal of Risk and Uncertainty, 30,* 241–259. http://dx.doi.org/10.1007/s11166-005-6563-7

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science, 18,* 105–110. http://dx.doi.org/10.1111/j.1467-9280.2007.01856.x

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review, 88,* 848–881.

Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science, 4,* 282–289. http://dx.doi.org/10.1214/ss/1177012493

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12,* 1062–1068. http://dx.doi.org/10.1038/nn.2342

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118,* 523–551. http://dx.doi.org/10.1037/a0024558

Grosskopf, B., Erev, I., & Yechiam, E. (2006). Foregone with the wind: Indirect payoff information and its implications for choice. *International Journal of Game Theory, 34,* 285–302. http://dx.doi.org/10.1007/s00182-006-0015-8

Hau, R., Pleskac, T. J., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making, 23,* 48–68. http://dx.doi.org/10.1002/bdm.665

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539. http://dx.doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115,* 225–237. http://dx.doi.org/10.1016/j.cognition.2009.12.009

Hoffrage, U., Weber, A., Hertwig, R., & Chase, V. M. (2003). How to keep children safe in traffic: Find the daredevils early. *Journal of Experimental Psychology: Applied, 9,* 249–260. http://dx.doi.org/10.1037/1076-898X.9.4.249

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., . . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8,* 75–84. http://dx.doi.org/10.1037/1076-898X.8.2.75

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Lewandowsky, S., Oberauer, K., Yang, L.-X., & Ecker, U. K. H. (2010). A working memory test battery for MATLAB. *Behavior Research Methods, 42,* 571–585. http://dx.doi.org/10.3758/BRM.42.2.571

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences, USA, 104,* 9493–9498. http://dx.doi.org/10.1073/pnas.0608842104

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 167–185. http://dx.doi.org/10.1037/0278-7393.34.1.167

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. http://dx.doi.org/10.2307/271063

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1355–1370. http://dx.doi.org/10.1037/0278-7393.32.6.1355

Rieskamp, J., Busemeyer, J. R., & Laine, T. (2003). How do people learn to allocate resources? Comparing two learning theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1066–1081. http://dx.doi.org/10.1037/0278-7393.29.6.1066

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

Schachter, S., Ouellette, R., Whittle, B., & Gerin, W. (1987). Effects of trend and of profit or loss on the tendency to sell stock. *Basic and Applied Social Psychology, 8,* 259–271. http://dx.doi.org/10.1207/s15324834basp0804_1

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames: A learning theory approach. *Journal of Economic Behavior & Organization, 7,* 47–70. http://dx.doi.org/10.1016/0167-2681(86)90021-1

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

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review, 112,* 862–880. http://dx.doi.org/10.1037/0033-295X.112.4.862

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale: Revised manual*. New York, NY: Psychological Corporation.

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review, 12,* 387–402. http://dx.doi.org/10.3758/BF03193783

Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making, 19,* 1–16. http://dx.doi.org/10.1002/bdm.509

Received December 21, 2012
Revision received August 1, 2014
Accepted August 12, 2014 ■