Hold It!

The Influence of Lingering Rewards on Choice Diversification and Persistence

Christin Schulze[1,2], Don van Ravenzwaaij[1,3], and Ben R. Newell[1]

1. University of New South Wales, Sydney, Australia
2. Max Planck Institute for Human Development, Berlin, Germany
3. University of Groningen, Netherlands

Author Note

Christin Schulze, Don van Ravenzwaaij, and Ben R. Newell, School of Psychology, University of New South Wales, Sydney, Australia.

Christin Schulze is now at the Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. Don van Ravenzwaaij is now at Department of Psychology, University of Groningen, Netherlands.

We thank Ehsan Arabzadeh, Fred Westbrook, and Justine Fam for valuable discussions about the application and scope of this choice paradigm with animal and human subjects and Susannah Goss for editorial assistance. This research was supported by Australian Research Council grants to Ben R. Newell (DP110100797; FT110100151).

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

Word Count (excl. title page, abstract, references, tables, and figure captions): 9612

Abstract

Learning to choose adaptively when faced with uncertain and variable outcomes is a central challenge for decision makers. This study examines repeated choice in dynamic probability learning tasks in which outcome probabilities changed either as a function of the choices participants made or independently of those choices. This presence/absence of sequential choice–outcome dependencies was implemented by manipulating a single task aspect between conditions: the retention/withdrawal of reward across individual choice trials. The study addresses how people adapt to these learning environments and to what extent they engage in two choice strategies often contrasted as paradigmatic examples of striking violation of versus nominal adherence to rational choice: diversification and persistent probability maximizing, respectively. Results show that decisions approached adaptive choice diversification *and* persistence when sufficient feedback was provided on the dynamic rules of the probabilistic environments. The findings of divergent behavior in the two environments indicate that diversified choices represented a response to the reward retention manipulation rather than to the mere variability of outcome probabilities. Choice in both environments was well accounted for by the generalized matching law, and computational modeling-based strategy analyses indicated that adaptive choice arose mainly from reliance on reinforcement learning strategies.

Keywords: diversification, probability matching, statistical independence, cognitive models, decision making

Hold It!

The Influence of Lingering Rewards on Choice Diversification and Persistence

"Skate to where the puck's going, not where it's been" (MacGregor, 1999, p. 20). This advice given to the young Wayne Gretzky by his father, Walter, undoubtedly contributed to a remarkable career in professional ice hockey. It illustrates a key characteristic of everyday decision making: Most choices are made against the backdrop of a constantly changing environment, and what may have been a good option today will not necessarily work tomorrow. Sports competitions are just one example of situations in which prior actions and future outcomes are not independent. For example, a recent move may decrease the likelihood of the same move being successful again simply because opponents can anticipate it; unexpected moves are more likely to be effective. In other words, successful choices on the playing field—and in everyday life—require anticipation and adaptation to an ever-changing world that is shaped by the decisions of those involved.

By contrast, consider a common laboratory choice paradigm that examines sequential decision making in a perfectly static environment. A fair, ten-sided die with seven green and three red sides is rolled numerous times, decision makers are asked to bet on a color for each roll, and correct bets are rewarded with fixed payoffs (e.g., Gal & Baron, 1996; James & Koehler, 2011; Peterson & Ulehla, 1965). The roll of a fair die is a paradigmatic example for statistical independence of a succession of events. The probability of rolling green on the current trial does not affect the probability of rolling red/green on subsequent trials, and the odds of success for each color remain stationary. Thus, payoffs are maximized by betting exclusively on the color with the higher reward likelihood—i.e., by *probability maximizing*. Maximization remains the best strategy even when outcome probabilities are not explicit. Many repeated choice paradigms are probability learning tasks with stationary and statistically independent outcome probabilities (see, e.g., Estes, 1964). Despite the apparent simplicity of the problem, people often need hundreds of choice trials and extensive feedback to eventually adopt and maintain a maximizing strategy (e.g., Newell, Koehler, James, Rakow, & van Ravenzwaaij, 2013; Newell & Rakow, 2007; Shanks,

Tunney, & McCarthy, 2002). Instead, they initially diversify their choices and bet on the dominant color on 70% of rolls and its alternative on the remaining 30%—i.e., they *probability match*—or they select the dominant color slightly more often than indicated by its outcome probability (over-matching; for reviews, see Koehler & James, 2014; Vulkan, 2000). As the outcomes are temporally independent, however, choice diversification results in inferior choice accuracies, and probability matching and its variants are considered violations of rational choice theory.

Although irrational in perfectly constant, transparent choice settings, the tendency to diversify may be a reasonable response to limited and uncertain information about the random nature of many repeated choice tasks. For instance, the behavior of an optimal Bayesian algorithm initialized with the erroneous—but ecologically plausible—belief that outcomes are temporally interdependent has been shown to converge on probability matching not probability maximizing (Green, Benson, Kersten, & Schrater, 2010). Additionally, choice diversification via probability matching has been found to occur at a global level partly because people engage in an exploratory strategy that is highly adaptive in most natural environments but futile under sequential independence: the search for patterns in the outcome sequence (e.g., Gaissmaier & Schooler, 2008; Peterson & Ulehla, 1965). If a pattern did exist, a rule-searching strategy that identified the pattern would be superior to static probability maximizing, and it has been shown that decision makers who probability matched in the absence of patterns were more likely to detect regularities in the outcome sequence when patterns were introduced (Gaissmaier & Schooler, 2008). Thus, choice diversification in static probability learning paradigms may be an over-learned response from frequently encountered dynamic real-world settings—that is, environments which are shaped by sequential choice–outcome dependencies and in which spreading of choice allocations can be beneficial.

Now consider another prominent method for studying choice, established in the context of operant conditioning: concurrent schedules of reinforcement. In common concurrent reinforcement procedures, an animal selects among two (or more) choice options and the ensuing reinforcement at each option depends

on, for instance, the passage of time (concurrent variable-interval schedules; e.g., Herrnstein, 1961), rates of prior choices of the selected option (concurrent variable-ratio schedules; e.g., Herrnstein & Loveland, 1975), or rates of prior choices of any option (dependent concurrent variable-ratio schedules; e.g., MacDonall, 1988). Once available, these reinforcers, or rewards, typically remain so until collected. Hence, concurrent reinforcement differs from probability learning procedures in that reward availability at each choice does not solely depend on a random process that delivers (or omits) a reward with a predetermined probability but is also influenced by choices made in the past. This arrangement more closely reflects features of natural choice environments (see, e.g., Ayton & Fischer, 2004). For instance, foraging animals are faced with the challenge that most food resources are exhaustible (what an animal consumes is gone), replenish independently of their exploitation (resources are deposited independently of an animal's feeding), and remain available until consumed or perished (potential retention of resources). Under these conditions, choices that have been successful in the past will not necessarily succeed in the future. Conversely, choices that were not successful in the past may have become more favorable since they were last visited.

Drawing a connection between these two prominent approaches to studying choice—probability learning and concurrent reinforcement—we contrast human sequential choice when reward availability changes either as a function of participants' choices or independently of those choices. The presence/absence of sequential choice–outcome dependencies was implemented by manipulating a single task aspect that distinguishes between probability learning and concurrent reinforcement procedures: reward hold, that is, the retention of a reward across trials (for similar approaches in the context of animal choice, see, e.g., Fam, Westbrook, & Arabzadeh, 2015; Jensen & Neuringer, 2008; MacDonall, 1988).

Under *reward hold*, rewards remained available after their initial scheduling, as in concurrent schedules of reinforcement. Specifically, we used a discrete-trial concurrent reinforcement procedure, in which choices of either of two options advanced the reinforcement schedules of both options and scheduled rewards remained available at each option until that option was next selected (see, e.g., Fam et

al., 2015; Jensen & Neuringer, 2008; Lau & Glimcher, 2005; MacDonall, 1988; Rothstein, Jensen, & Neuringer, 2008; Rutledge et al., 2009; Serences, 2008). This procedure introduces sequential dependencies between reinforcement probabilities and successive choices: The probability of being rewarded increases with time spent not choosing an option, because rewards scheduled to occur are held until collected. The top part of Figure 1 illustrates this relationship by depicting how reward probabilities—initially set at .70 and .30—change as a function of four hypothetical choice sequences under reward hold (see the Method section for details on how we computed this). Each hypothetical choice is denoted by the letter H or L depending on whether the option with the initially higher or lower outcome probability is selected. The ensuing reward probabilities at option H are shown by solid lines; those at option L by dashed lines. With each choice, the likelihood of obtaining a reward at the unchosen option increases as shown, and a reward maximizing strategy involves switching between options based on these changing outcome probabilities. This strategy is exemplified by the first hypothetical choice sequence in the top half of Figure 1. In this example, a participant chooses option H on the first three trials but switches to option L on the fourth trial because the probability of obtaining a reward from option L now exceeds the probability of obtaining a reward from the H (note that the dashed line, the reward probability at option L, is above the solid line, the reward probability at option H, on the fourth trial). By contrast, continuous exploitation of a single option—as illustrated by the second and fourth hypothetical choice sequences in Figure 1—ignores these probability changes and leads to lower expected reward. Thus, under reward hold, choice diversification based on choice–outcome dependencies is superior to choice persistence.

[insert Figure 1 here]

Under *no hold*, rewards were available only when scheduled to occur, as in standard probability learning paradigms. To keep all other task aspects constant between these conditions, we had reward probabilities change dynamically in a similar way as under reward hold but unaffected by participants' prior decisions. The bottom part of Figure 1 illustrates how we derived these dynamic outcome

probability changes under no hold from actually observed probabilistic changes under hold (see the Method section for details). This procedure reinstated the classic advantage of probability maximizing: Although outcome probabilities varied throughout the experiment, one option almost always had a higher likelihood of delivering a reward. The central goal of this approach was to directly compare whether people learn to adaptively engage in either choice diversification or persistent maximization—two choice strategies often contrasted as paradigmatic examples of striking violation of versus nominal adherence to rational choice—in dynamic sequential choice tasks that link two prominent approaches to the study of decision making.

We are not the first to contrast choice under probability learning and concurrent reinforcement. Only a small distance separates the two paradigms at the level of procedure, and the choice phenomena typically observed within each are conceptually related (see Herrnstein, 1970; Herrnstein & Loveland, 1975). However, to our knowledge, this is the first investigation that contrasts human rather than non-human animal decision making in these two paradigms and that applies this particular manipulation to render the paradigms comparable on all but the reward hold dimension. Research with animal subjects has suggested that choice behavior in both paradigms may be characterized by a common set of principles: the *matching law* (Herrnstein, 1970; Herrnstein & Loveland, 1975). Herrnstein's (1961) matching law states that the proportion of choices of an option will equal the proportion of reinforcement obtained from that option. Although terminologically similar, the matching law and probability matching make different predictions in standard probability learning paradigms. Probability matching predicts that choice rates will be proportional to the options' programmed outcome probabilities; the matching law predicts that choice rates will be proportional to the average reinforcement obtained from each choice alternative. Obtained and programmed reinforcement are not equivalent under probability learning; the former depends on actual choice distributions and the latter does not. In a discrete-trial probability learning task, in which two choice alternatives yield reward with differing probability, the predictions of the matching law will be met only if one of the alternatives is selected exclusively (Herrnstein &

Loveland, 1975). Thus, the matching law predicts asymptotic probability maximizing—not matching—in probability learning paradigms. Human subjects' choice behavior early in a probability learning paradigm typically resembles probability matching (see Vulkan, 2000). Yet after extensive training, asymptotic choice behavior closely approximating probability maximizing has been observed (Shanks et al., 2002)—as is predicted by the matching law. Moreover, generalized to

$$\log\left(\frac{B_1}{B_2}\right) = a\log\left(\frac{R_1}{R_2}\right) + \log b, \quad (1)$$

where B_i denotes the response frequency to alternative i and R_i denotes the reward obtained from that option, the *generalized matching law* accounts for systematic deviations from strict matching under various conditions of reinforcement (Baum, 1974). Specifically, the slope of the generalized function captures the sensitivity of choice to changes in the reinforcement ratio; the intercept expresses bias not accounted for by reinforcement rates (Baum, 1974, Baum, 1979). The generalized matching law provides a good description of behavior in various species (see McDowell, 2013) and, applied to real-world decision settings, has provided insights into human behavior in numerous social and clinical contexts (see, e.g., Borrero et al., 2007; Borrero & Vollmer, 2002; Reed, Critchfield, & Martens, 2006).

Studies contrasting reward hold and standard probability learning environments have found that the generalized matching law provides a good description of the choice behavior of different animal species in both these situations (Fam et al., 2015; Jensen & Neuringer, 2008; MacDonall, 1988). Pigeons' choices, for instance, are captured by the generalized matching law when evaluated along a continuum of varying degrees of reward hold, the ends of which are marked by the two choice paradigms of interest here (Jensen & Neuringer, 2008). Thus, by applying a similar approach to human choice, our study extends this line of research and aims to determine whether humans learn to diversify and to persist when probability learning and concurrent reinforcement are linked within a single paradigm characterized by the presence/absence of sequential choice–outcome dependencies.

A second goal of our study was to examine *how* people might accomplish learning in choice environments that require either diversification or persistence. To this end, we manipulated the feedback available during the task. We contrasted partial feedback about obtained reward at the chosen option only with full feedback about both obtained and forgone reward. This additional feedback about forgone outcomes was hypothesized to aid responding in both environments—but potentially to a higher degree under reward hold than under no hold. Under no hold, participants need to detect the absence of choice–outcome dependencies. Under hold, by contrast, they need to learn about the specifics of existing choice–outcome dependencies, which may depend on the availability of full outcome feedback. Additionally, we applied a computational modeling-based approach to examine the learning and choice mechanisms adopted by decision makers in our paradigm. After presenting the behavioral results, we describe the computational models of learning and heuristic decision making and evaluate them in light of participants' responses.

Method

Participants. Ninety-two (47 female) undergraduate students from the University of New South Wales with a mean age of 19.46 years (*SD* = 2.11 years) participated in exchange for course credit and a small performance-based payment (earnings ranged from AU$3.25 to AU$3.90; 1 AU$ ≈ 1 US$ at the time of the experiment). The experiment was approved by the ethics committee of the University of New South Wales.

Design and procedure. Participants completed a computer-based sequential binary choice task with 500 trials. We factorially crossed two between-subjects factors: the presence of reward retention across trials (reward hold vs. no hold) and the type of outcome feedback that participants received (partial vs. full feedback).[1] Under partial feedback, participants were informed about obtained payoffs only;

[1] Originally, we had planned to contrast these two feedback types to an additional between-subjects condition in which outcome information was manipulated via a cover story that hinted at either the presence or absence of reward hold. Due to

under full feedback, they received feedback about payoffs for the option they chose (obtained) as well as for the option they did not choose (forgone). To derive probability changes for conditions under which reward hold was absent from conditions under which it was present (see *Reward hold manipulation* section), we implemented the following data collection procedure: The first 15 participants in each of the four conditions were run sequentially in the following order: partial feedback under hold, partial feedback under no hold, full feedback under hold, and full feedback under no hold. The remaining eight participants in each condition (32 in total) were allocated randomly. The entire process of data collection spanned a period of approximately one month.

Participants were informed that they could earn performance-based payoffs in the choice task and were encouraged to attempt to earn as much money as possible. The choice task was segmented into five blocks of 100 trials each. After every block, a short motivational pop-up message encouraged participants to carry on and allowed for a self-paced pause in the experiment. Following the choice task and display of total earnings, all participants were asked to complete a short questionnaire assessing their understanding of the underlying probability structure.[2] Finally, participants were paid in cash according to the total amount of points they obtained in the experiment.

Sequential choice task. We used a standard probability learning paradigm involving repeated decisions between two unmarked buttons displayed on a computer screen over 500 choice trials (see, e.g., Erev & Barron, 2005). Two discrete outcomes were possible at both options: ten (reward) or zero points (no reward) were scheduled with initial probabilities of $p(\mathrm{H}) = .70$ (high option) and $p(\mathrm{L}) = .30$ (low option)—randomized across participants for left and right choice buttons. The reward probabilities of the options were statistically independent, such that either, neither, or both alternatives could yield a reward

concerns regarding the effectiveness of this manipulation in highlighting the desired aspects of the task environment, however, we do not consider these data further.

[2] One question asked participants to estimate the probability of being rewarded at each option. We found no significant differences in participants' probability estimates for the two options across experimental conditions (all $ps \geq .129$ and all $BFs \leq 0.78$). That is, the yoking procedure we applied between the hold and no hold environments effectively ensured that participants experienced similar reward rates. The data obtained from the post-task questionnaire were therefore not considered further.

on any given trial. Reward distributions across all choice trials were generated randomly to reflect these probabilities prior to the first trial for each participant. Following each choice, the available points (zero or ten) were revealed on either the selected button only (partial feedback) or both the selected and the unselected button simultaneously (full feedback); the total score was continuously updated and displayed on the screen. Obtained points were converted to cash payments at the end of the experiment at a rate of 10 points = 1 cent and participants were encouraged to attempt to earn as many points as possible.

Reward hold manipulation. Under reward hold, any points present at a choice option remained available for subsequent trials until collected. However, points being held could not accumulate, that is, there was never more than a single reward available at any option. With this reward hold manipulation, individual trial outcomes depended on programmed outcome probabilities as well as on participants' prior choices: The probability of being rewarded at either option increased with choices of the other option, because rewards scheduled to occur were held until collected. Specifically, the probability of obtaining a reward from option *i* following *n* choices of the other option (here denoted as $r(i)_n$) was determined by the programmed outcome probability, $p(i)$, and the number of choices of the other option (*n*) since option *i* was last selected (for similar calculations, see Jensen & Neuringer, 2008), such that

$$r(i)_n = 1 - \big(1 - p(i)\big)^{n+1}. \quad (2)$$

The top half of Figure 1 illustrates this relationship by depicting probability changes for the high and low option as a function of four hypothetical choice sequences. The lines for the second hypothetical choice sequence, for example, indicate that the reward probability at the low option surpasses the programmed *p*(H) of .70 after three consecutive choices of the high option, as given by Equation 2, $r(L)_3 = 1 - (1 - .30)^{3+1} = .7599$. Thus, with this particular concurrent reinforcement procedure, neither alternative represents a superior choice throughout. Instead, both options need to be selected in accordance with changing reward probabilities in order to maximize payoffs. Such a diversification strategy is illustrated by the first hypothetical choice sequence depicted in the top part of Figure 1.

Average reward proportions expected from different diversification strategies can readily be calculated on the basis of Equation 2. Table 1 compares the rewards expected for various diversifying choice sequences with the success rate of static exploitation of either option. Each cell in the table shows the average reward proportion expected for a choice sequence consisting of y (row number) consecutive choices of the high and x (column number) consecutive choices of the low option. For example, alternating between the high and low choice option at a rate of 3:1 as depicted by the first hypothetical choice sequence in the top part of Figure 1—and thereby maximizing *local* reward—yields reward on an average of .767 choice trials. This value is obtained by averaging across the expected reward from all four choices in this sequence: $r(L)_3 = .7599$ (selecting the low option following three choices of high), $r(H)_1 = .91$ (selecting the high option following one choice of low), and two times $r(H)_0 = .70$ (programmed outcome probability for two consecutive choices of the high option). As Table 1 shows, numerous diversification strategies (shaded in gray) provide higher expected reward than probability maximizing—a reward-maximizing strategy in stationary probability learning paradigms—which is confined by the reward rate of the continuously exploited "high" option (i.e., .70). For subsequent analyses, we placed particular emphasis on those sequences that provided maximum advantage over static probability maximizing (i.e., average reward proportion > .76; see the dark shaded cells in Table 1). Note that all four of these sequences involve only a single choice of the low option. That is, although reward retention makes it beneficial to switch recurrently to the low option, this option should never be chosen consecutively. Such deterministic cycles of a fixed number of choices of the richer alternative followed by a single choice of the leaner alternative also maximize reward in other discrete-trial concurrent reinforcement procedures, in which the scheduling of rewards is contingent on the passage of time rather than on prior choice rates (see Houston & McNamara, 1981; Staddon, Hinson, & Kram, 1981).

[insert Table 1 here]

Under no hold, points were not transferable across trials but could only be collected when scheduled to occur. To keep all features except the reward hold manipulation constant between the two

conditions, we derived outcome probabilities from the probabilistic changes that occurred under reward hold as a function of participants' choices. Specifically, we averaged the trial-by-trial probabilities of obtaining a reward from each option as given by Equation 2 across the first 15 participants in each experimental condition under reward hold, which produced dynamically fluctuating high and low reward probabilities. The bottom part of Figure 1 illustrates this procedure. The mean high and low outcome probabilities across all choice trials derived with this method were .76 and .63 for the partial feedback and .76 and .65 for the full feedback condition. Although the outcome probabilities fluctuated around their means, one option had a higher reward probability than the other on the majority of trials (94% and 89% of trials in the partial and full feedback condition, respectively). Note that this procedure focused on a subset of participants (15 out of 23) to keep the experienced outcome variability under no hold sufficiently high. This approach created a dynamic decision environment analogous to that experienced by participants under reward hold. The only important difference was that, irrespective of participants' decisions, one choice option was superior on most trials and should therefore be selected exclusively throughout the task. Consequently, under no hold, static probability maximizing represented the optimal strategy.

Behavioral Results

In addition to conventional methods of hypothesis testing, we conducted Bayesian analyses, for which we report Bayes factors (*BF*) quantifying the strength of evidence in favor of the presence of an effect (see Jamil et al., 2016; Rouder, Morey, Speckman, & Province, 2012).[3] A *BF* of 10, for instance, suggests that the data are 10 times more likely to have occurred under the model assuming the relevant

[3] All Bayesian analyses were carried out in R. For Bayesian ANOVAs and contingency tables we used the anovaBF and contingencyTableBF functions included in the BayesFactor package (v.0.9.12-2; Morey & Rouder, 2015) with their respective default settings, with one exception: For the anovaBF function, the number of Monte Carlo samples used to estimate *BF*s was increased to 50,000. For Bayesian ANOVAs, we computed *BF*s by contrasting the performance of a model including the relevant effect to one omitting (only) that effect. All models were constructed hierarchically, such that the presence of an interaction term always involved the presence of all lower order interactions and main effects involving the components of that higher order interaction. For Bayesian contingency tables, *BF*s in favor of the alternative hypothesis are reported.

effect than under a model omitting it, whereas a *BF* of 0.10 indicates that the data are 10 times more likely to have occurred under the model omitting the relevant effect than under a model including it.

Aggregate choice behavior. Recall that under reward hold, both choice options needed to be selected in line with the changing reward probabilities to reap maximum rewards. Choice sequences with a clear advantage over static probability maximizing (e.g., average reward proportion ≥ .76 instead of .70; see Method) are characterized by a series of (3–6) choices of the high option followed by a single choice of the low option. Thus, a range of diversifying choice proportions (e.g., .75–.86 choices of high) could potentially yield high reward in the hold environment. Under no hold, a single choice option was superior throughout and should therefore be selected exclusively to maximize reward (i.e., 1.00 choices of high). Figure 2 displays the mean proportion of choices of the high option for each block of 100 trials averaged across participants in each experimental condition. A 2 (reward hold) × 2 (feedback type) × 5 (block) mixed model ANOVA with choice proportion as the dependent variable showed a significant main effect of probability learning across trial blocks, $F(3.44, 302.45) = 7.66$, $p < .001$, $\eta_p^2 = .080$, $BF = 2819.00$; average choice proportions more closely resembled random guesses in the first 100 trials than in the last block of 100 trials.[4] We also found a significant main effect of feedback type, $F(1, 88) = 5.54$, $p = .021$, $\eta_p^2 = .059$, $BF = 2.76$; participants who received full outcome feedback selected the high option more frequently ($M = .74$, across blocks) than did those who received partial feedback ($M = .67$, across blocks). There was no significant effect of reward retention, $F(1, 88) = 0.24$, $p = .625$, $\eta_p^2 = .003$, $BF = 0.32$, and none of the interactions reached statistical significance (all ps ≥ .252 and all BFs ≤ 0.62). Thus, *on average*, participants in both choice environments had similar choice proportions.

[insert Figure 2 here]

Individual choice behavior. Figure 3 displays the distributions of individual participants' proportion of high-option choices and reveals a different pattern of results: Individuals' choice strategies

[4] For all conventional ANOVAs, degrees of freedom were corrected using the Greenhouse and Geisser (1959) coefficient when the sphericity assumption had been violated.

under reward hold were strikingly different from those employed under no hold. Under hold, participants' choice distributions approached a unimodal peak around effective diversification; under no hold, distributions were rather bimodal, with some participants maximizing accurately and others showing indifference between the two options (i.e., random responding). Moreover, nearly all participants in the hold environment (22 out of 23 in each feedback type condition) responded in a manner inconsistent with static probability maximizing, which is an inferior strategy in this environment.

[insert Figure 3 here]

To evaluate strategy selection toward the end of learning, we classified participants' response proportions in the final trial block as (a) probability maximizing (choosing the high option in no less than 95% of trials; see, e.g., Newell & Rakow, 2007), (b) potentially effective diversification (choosing the high option in 75%–86% of trials),[5] or (c) other strategy use. Table 2 summarizes the association between individual participants' strategy use and the hold manipulation within and across feedback conditions. Across feedback conditions, the odds of following a potentially effective diversification strategy rather than any other strategy were 6.94 times higher under reward hold than under no hold. Similarly, across feedback conditions, the odds of probability maximizing were 9.63 times higher in a no hold environment than in a hold environment. Thus, comparable (and substantial) numbers of participants in both environments applied adequate choice strategies.

[insert Table 2 here]

[5] Strictly speaking, in the full feedback hold environment, effective diversification as defined via these choice proportions is not necessarily optimal. This is because information about forgone outcomes allows participants to discern when exactly it is time to collect a reward at the low option, and cyclic checking as defined by these choice proportions becomes unnecessary. Thus, in this environment, it would be optimal to continuously select the high option until the feedback indicates that a reward is available at the low option, at which time this option should be selected immediately (and only once). None of our participants in the full feedback hold condition strictly followed this strategy during the final trial block and only few approached it. Specifically, the choices of three participants were consistent with this strategy on at least 90% of trials during the final trial block. The choice proportions of two of these three participants fell within the range of what we defined as effective diversification (choosing the high option in 75%–86% of trials), whereas the third participant selected the high option in 74% of trials during the final trial block.

Choice in relation to obtained reward. The matching law predicts that choice rates will be proportional to the average reward obtained from each choice alternative. Figure 4 shows log response ratios as a function of log obtained reward ratios for each participant in each experimental condition. We used least squares linear regression to fit the logarithmic form of the generalized matching law (Equation 1) to participants' choice data averaged across all trials. The estimated parameters from these fits and the regression lines are displayed in Figure 4. We found close approximations of a perfect matching relationship in each experimental condition, but some slight deviations are noteworthy. Under no hold, regression slopes smaller than one indicate that participants showed a slight tendency to "undermatch," that is, choice allocation was less extreme than predicted by the matching law. This is consistent with the notion that perfect matching in a discrete-trial probability learning paradigm can be achieved only by exclusively selecting one of the alternatives (Herrnstein & Loveland, 1975)—a behavior that not all participants in the no hold environment showed (see Figure 3). Moreover, as indicated by negative *y*-intercepts, participants in all conditions (but more so under reward hold) showed a slight bias toward the low option that was unrelated to differences in the reward rate.

[insert Figure 4 here]

Overall, the generalized matching law was able to account for average choice behavior across trials of participants in both choice environments, explaining at least 90% of the variance in the data for each condition (see Figure 4).

Interim Discussion

In a dynamic choice environment in which outcome probabilities changed contingent on prior decisions because reward was held until collected, choice behavior was found to remain variable. That is, under reward hold, most participants responded in a manner inconsistent with inferior probability maximizing and many implemented effective choice diversification (approximately one third of participants). Under no hold, similar numbers of participants behaved adaptively, that is, despite

experiencing variable and closely proximate outcome probabilities, a substantial subset of participants identified probability maximizing as an effective strategy. In both environments, effective strategy use was aided by trial-by-trial feedback about forgone outcomes to a similar extent. That is, noticing the absence of choice–outcome dependencies and identifying existing choice–outcome dependencies was facilitated by the availability of full outcome feedback to a similar degree. Our findings of different individual choice distributions in the two environments indicate that diversified choices under reward hold represented a response to the reward retention manipulation rather than incomplete learning. It did not seem to be the case that participants under reward hold switched among the options because they were still *exploring* how best to respond but because they had *learned* that diversification is beneficial. If they were still exploring, we would have expected to see similar individual choice proportions in the no hold environment. This is because, under no hold, the probability of reward also changed dynamically—and crucially, the difference between reward probabilities at the two choice options mirrored the probabilities experienced by participants under hold—but outcomes were no longer dependent on prior choices. In this environment, however, participants showed a stronger tendency to probability maximize, which was adaptive in this context.

Average choice in both environments was well described by the generalized matching law and, aside from a slight bias in choice toward the low option, we found an almost perfect matching relationship in each experimental condition. This bias may indicate higher valuation of rewards obtained from the leaner alternative (Baum, 1974), which is in line with findings from probability learning paradigms suggesting that persistent probability matching may arise from the higher satisfaction associated with correctly guessing a rare rather than a common event (see, e.g., Goodnow, 1955). The matching law, however, is silent with respect to the trial-by-trial characteristics underlying average choice across trials. Thus, we now turn to a computational modeling-based approach to illuminate trial-by-trial response dynamics, learning, and strategy selection in our task.

Model-Based Strategy Analysis

How did people learn to choose effectively under reward hold and no hold when choice–outcome dependencies were either present or absent? Which strategies best account for the adequate choice policies that were applied by some participants in both environments? What influence did the differing types of feedback have on strategy selection? To address these questions, we examined various cognitive models of choice potentially underlying participants' choice behavior in the two environments.

Overview of Strategy Models

We examined three simple choice strategy models, including the win-stay lose-shift (WSLS) heuristic (e.g., Nowak & Sigmund, 1993) and reinforcement-based learning (Sutton & Barto, 1998). WSLS and reinforcement-based learning rest on the assumption that choices are made in accordance with previous reward experience and have been proposed to describe behavior in related decision tasks (e.g., Busemeyer & Stout, 2002; Gaissmaier & Schooler, 2008; Gureckis & Love, 2009; Otto, Taylor, & Markman, 2011; Worthy, Otto, & Maddox, 2012). Additionally, we developed and evaluated a simple reward-independent strategy model that derives choice probabilities from assumed outcome regularities, irrespective of prior reinforcement. All strategy models were compared with a baseline model that fitted a constant and statistically independent choice probability of selecting the initially more probable option to each participant's choices across trials. The full model specifications, the parameter estimation method, and the model comparison procedure are described in detail in the Appendix.

Win-stay lose-shift. This simple rule-based strategy assumes that a decision maker continues to choose an option following a reward ("win-stay"), but shifts to the other option following the absence of a reward ("lose-shift"). Here, we consider a probabilistic modification of this deterministic rule that allows free variation of the probability of repeating a choice following a success and the probability of switching following a failure. WSLS requires no memory because subsequent choice probabilities are computed

solely on the basis of current outcomes. This strategy has been shown to contribute to choice diversification in binary choice tasks (e.g., Gaissmaier & Schooler, 2008; Otto et al., 2011).

Reinforcement learning. Models of reinforcement learning (see, e.g., Sutton & Barto, 1998) allow for reliance on longer windows of prior reward experience and have been successful in describing choice in similar contexts (see, e.g., Busemeyer & Stout, 2002; Gureckis & Love, 2009; Schulze, van Ravenzwaaij, & Newell, 2015). We consider a simple learning model that assumes that a decision maker gradually updates propensities toward each choice alternative based on the observed discrepancy between expected and actual monetary rewards, weighted by a learning rate parameter. High values of this learning rate parameter indicate strong reliance on recent outcomes; low values indicate integration of larger windows of past experience. On the basis of these preferences, the decision maker selects an option on each trial with a degree of precision that is determined by a sensitivity parameter. Low values of this sensitivity parameter indicate low choice precision in terms of more random guessing; high values indicate strictly deterministic selection of the preferred alternative.

Choice pattern. We developed a simple strategy model which encompasses a range of choice rules that vary in the extent to which they are deterministic and which can account for systematic patterns in participants' responses. Evaluating the prevalence and form of patterned responding is instructive for both environments. In the reward hold environment, maximum rewards could be obtained by following highly structured patterned sequences (e.g., four choices of high, one of low, four of high, etc.; see Table 1). In the no hold environment, outcomes were statistically independent across trials, but participants in similar probability learning tasks regularly engage in pattern search despite the statistical independence of outcomes (e.g., Gaissmaier & Schooler, 2008).

The choice pattern model assumes that decision makers have preferred successive choice run lengths for both options, which describe the regularity of the choice pattern employed. Specifically, these preferred choice run lengths determine the probability with which a decision maker shifts to the other option after a run of successive choices of one option. Additionally, the model allows these shift

probabilities to change either rapidly or gradually in response to the choices made. That is, varying degrees of precision in choice rule adherence can be implemented. This model thus parameterizes three key characteristics of participants' choice sequences that we can examine directly: (a) preferred choice run length for the high option, (b) preferred choice run length for the low option, and (c) precision of choice rule adherence.

Modeling Results

Strategy classification. The three choice strategy models and the simple baseline model were fit to individual participants' choices using a maximum likelihood approach. For each participant, the models were evaluated using Akaike's Information Criterion (AIC; Akaike, 1974), and each participant was classified as using the strategy with the lowest AIC model weight (see the Appendix for a detailed description of the parameter estimation and model evaluation procedure). Figure 5 shows the percentage of participants in each experimental condition for whom each model provided the best fit. Across all conditions, only a small subset of participants was best described by the baseline model (8%), which assumes a constant rate of responding throughout the task. This result is consistent with the learning effects we observed in the experiment, which are illustrated by the upward trend in average choice proportions shown in Figure 2. The choice behavior of most participants was best described by either a reinforcement learning mechanism (41%) or a simple WSLS heuristic (38%), indicating that participants in both environments were sensitive to the scheduling of rewards and adjusted their behavior accordingly. By contrast, only few participants were classified as using a reward-insensitive choice pattern (13%). Of these participants best described by the choice pattern model, most experienced no hold and full outcome feedback; in fact, for this experimental condition, more participants were best fit by the pattern model than by any other model. Model-based strategy classifications were related to the use of adequate choice policies in both environments. Most participants whose choice proportions in the final trial block were identified as adequate in the respective environment (i.e., probability maximizing under no hold and effective diversification under hold; see *Individual choice behavior* section) were classified as users of a

reinforcement learning strategy (47% under hold and 57% under no hold). In other words, effective choice in both environments mainly arose from the application of reinforcement learning strategies rather than memory-free heuristics or reward-independent choice patterns.

[insert Figure 5 here]

Parameter analysis. How did the different choice environments and feedback conditions impact the choice mechanisms captured by the three strategy models? Table 3 summarizes average parameter estimates for each evaluated strategy model (excluding those of the baseline model, which reflect the average choice proportions illustrated in Figure 2) and experimental condition, which were subjected to 2 (reward hold) × 2 (feedback type) ANOVAs. For the WSLS model, we found significant effects of the reward hold manipulation on both the probability of staying after a win, $F(1, 88) = 7.94$, $p = .006$, $\eta_p^2 = .083$, $BF = 6.92$, and the probability of switching after a loss, $F(1, 88) = 45.74$, $p < .001$, $\eta_p^2 = .342$, $BF = 8.25 \times 10^6$. Participants under reward hold were less likely to persevere at an option after a win but more likely to switch after a loss (see Table 3), indicating a stronger tendency to diversify choice, which was adaptive in this environment. Additionally, the conventional method of null hypothesis significance testing suggested an effect of feedback type on the probability of staying after a win but the Bayesian evidence was ambiguous, $F(1, 88) = 4.40$, $p = .039$, $\eta_p^2 = .048$, $BF = 1.51$. No other effects on parameters of the WSLS model were statistically significant (all ps $\geq .053$ and all BFs ≤ 1.15).

[insert Table 3 here]

For the reinforcement learning model, we found a significant effect of feedback type on the learning rate, $F(1, 88) = 12.45$, $p = .001$, $\eta_p^2 = .124$, $BF = 39.39$; when only partial feedback was available, participants integrated shorter windows of past experience (see Table 3), consistent with the higher need for exploration when no information about forgone outcomes is available. Additionally, null hypothesis significance testing suggested an effect of feedback type on the choice sensitivity parameter but the Bayesian evidence was ambiguous, $F(1, 88) = 4.59$, $p = .035$, $\eta_p^2 = .050$, $BF = 1.61$. We found no significant effects of reward hold on estimated reinforcement learning parameters (for all main effects and

interactions, *p*s ≥ .066 and *BF*s ≤ 1.15) but the direction of trends was consistent with the requirements of adaptive choice in both conditions. That is, under no hold, participants integrated slightly longer windows of past information and more deterministically exploited options they had learned to be profitable, whereas under hold, participants were more sensitive to momentary changes in the options' profitability and were more explorative in their action selection (see Table 3).

For the choice pattern model, we found significant effects of the hold manipulation on the preferred choice run length for the high option, $F(1, 88) = 10.45$, $p = .002$, $\eta_p^2 = .106$, $BF = 18.24$, the preferred choice run length for the low option, $F(1, 88) = 22.07$, $p < .001$, $\eta_p^2 = .200$, $BF = 1982.58$, and the precision of choice rule adherence, $F(1, 88) = 6.42$, $p = .013$, $\eta_p^2 = .068$, $BF = 3.73$. Participants under hold preferred shorter choice run lengths for both options but adhered to their preferred choice rules slightly *more* consistently than participants under no hold (see Table 3). Thus, although choice rule precision was generally very low, participants in the hold environment appeared to be slightly more systematic in following choice rules, which can be adaptive when reward availability changes systematically as a function of prior decisions. No other effects on parameters of the choice pattern model were statistically significant (all *p*s ≥ .135 and all *BF*s ≤ 0.75).

Discussion

Computational modeling-based analyses revealed large individual differences in choice strategy use during the experiment. Overall, reinforcement learning mechanisms were best able to account for most participants' choice behavior and were positively related to response proportions that signify adequate strategies under reward hold (effective diversification) and no hold (probability maximizing). By contrast, only few participants, most of whom responded under no hold, were classified as using a reward-independent choice pattern. Nevertheless, this latter result indicates that persistent choice under no hold *can* be captured by a reward-insensitive choice mechanism. That is, in the dynamic no hold environment, in which reward probabilities were proximate and variable, successful choice required participants to persist with the option that was more profitable overall, despite experiencing occasional

losses and having to disregard rewards at the other option. Indeed, previous research on stationary probability learning paradigms has suggested that a likely reason for many people's failure to adopt a probability maximizing strategy in these contexts is their unwillingness to accept the guaranteed loss rate associated with maximization (e.g., Arkes, Dawes, & Christensen, 1986).

Parameter analyses revealed that the underlying choice processes were largely consistent with the distinct demands of each experimental condition. Specifically, in the absence of full feedback about forgone outcomes, we found evidence for more explorative choice processes, and under reward hold, parameter estimates indicated stronger tendencies toward diversified choice (e.g., lower preferred choice run lengths for both options, and higher switch rates regardless of whether or not the previous action was rewarding) but also toward more precision in adhering to a particular choice pattern. In sum, the computational modeling-based approach illuminated the specific processes underlying adaptive choice behavior in our experiment and highlighted participants' sensitivity to the differential demands inherent to each combination of environmental factors examined.

General Discussion

Our study aimed to draw a connection between two prominent approaches to the study of choice, probability learning and concurrent reinforcement, by manipulating a pivotal task aspect that distinguishes between these paradigms: reward hold. We examined whether people learned to adaptively engage in either choice diversification or persistent maximization—two strategies frequently juxtaposed as examples of striking violation of versus nominal adherence to rational choice—when the manipulation of reward hold made outcomes either dependent on or independent of prior decisions. We found that choices both diversified and specialized when they should. That is, adaptive choice diversification and persistence manifested to similar extents. The generalized matching law captured these differing trends in choice behavior: on the one hand, asymptotic diversification in accord with obtained reward under hold; on the other hand, asymptotic probability maximizing under no hold. These findings are in line with previous research showing that the generalized matching law provides a good description of the choice

behavior of different animal species in both reward hold and standard probability learning environments (Fam et al., 2015; Jensen & Neuringer, 2008; MacDonall, 1988). Our study thus contributes to linking matching law and probability learning conceptually in the context of repeated human choice.

In addition to showing that the matching law is able to account for average choice under reward hold and no hold, we examined how different environmental characteristics led approximately one third of participants toward effective choice in each context. We found that the adoption of both adaptive choice diversification and persistence was facilitated to a similar extent by meaningful trial-by-trial outcome feedback. In this respect, our findings echo previous claims (e.g., Newell & Rakow, 2007; Shanks et al., 2002) that, if a learning environment is appropriately structured, many people can indeed learn to favor probability maximizing when it is optimal. Yet trial-by-trial feedback about forgone reward is no guarantee for improved performance when people make decisions based on repeated outcome experience. For instance, in stationary experience-based choice paradigms, information about forgone outcomes tends to intensify the underweighting of rare events (e.g., Yechiam & Busemeyer, 2006; Yechiam, Rakow, & Newell, 2015) and, in dynamic choice tasks, forgone feedback has been found to hinder people from maximizing long-term reward (Otto & Love, 2010). We, too, found that even when full feedback was given, not everyone learned to choose adaptively in our task; rather, we observed large interindividual differences in strategy selection.

Interindividual differences in learning outcomes in analogous but stationary choice tasks have been attributed to individual differences in cognitive capacity (Rakow, Newell, & Zougkou, 2010; West & Stanovich, 2003) and cognitive thinking styles (Koehler & James, 2010). Specifically, people with high cognitive reasoning abilities and strong deliberative cognitive reflection tendencies were found to be more likely to probability maximize in static choice contexts (Koehler & James, 2010; West & Stanovich, 2003). Moreover, people's decisions have been shown to improve when individual cognitive deficits are mitigated by raising strategy availability (Koehler & James, 2010; Newell et al., 2013), providing extensive feedback (Shanks et al., 2002), or enhancing cognitive processing in group choice contexts

(Schulze & Newell, 2016). We did not include individual difference measures in our study but our computational modeling-based analyses indicated that participants used different cognitive mechanisms to implement effective choice strategies. Previous research has found that, in static probability learning tasks, choice diversification strategies can arise from people applying simple cognitive processes that consider only the most recent outcome, or integrating a longer window of past outcomes depending on whether or not executive resources are taxed by additional task demands (Otto et al., 2011). Additionally, in discrete-trial concurrent reinforcement tasks similar to those we used in the hold environment, monkeys' choices have been found to be consistent with reinforcement learning (Lau & Glimcher, 2005). We found that effective diversification and persistence also arose mainly from reliance on reinforcement learning strategies.

Recall the challenges faced by Wayne Gretzky during an ice hockey game (or indeed by virtually anyone in day-to-day life). In a dynamic hockey game (world) that is shaped by players' (people's) actions, a recent good move (choice) will not necessarily work the next time. Winning (on the field and in life) requires effective anticipation and adaptation to constantly changing conditions. Accordingly, we observed that many participants learned to diversify when the learning environment was appropriately and ecologically plausibly structured. In this respect, our results contribute to a growing body of research demonstrating that people can learn to adaptively diversify their choices in ecologically valid contexts. For example, relating back to the choice diversification phenomenon introduced at the beginning of the article—probability matching—Gaissmaier and Schooler (2008) found that probability matchers were more likely to identify patterns in the outcome sequence and matched adaptively as a by-product of pattern search and adherence. Similarly, when the presence of a computerized opponent gives choice diversification an advantage over static maximization, adaptive human probability matching has been shown to occur under competition for available resources (Schulze et al., 2015).

Moreover, our findings mesh with studies that emphasize the importance of considering plausible assumptions that rational learners may hold when their behavioral outcomes are "irrational." As discussed

in Ayton and Fischer (2004), sequential outcome dependencies may be an environmental feature that people learn to expect on the basis of everyday experience. In an experimental choice task characterized by statistical independence, this assumption can result in ill-advised diversification (Green et al., 2010). Findings of optimal learning agents—typically Bayesian model-based—that hold ecologically plausible beliefs achieving suboptimal outcomes have also been demonstrated in connection with over-exploration (Navarro & Newell, 2014) and melioration (Sims, Neth, Jacobs, & Gray, 2013) in related paradigms. A similar argument has been used to explain another diversification strategy that occurs in some species without initial reinforcement: spontaneous alternation. When faced with two choice alternatives that never (or always) deliver a reward, foraging animals such as rodents, cats, and chickens sometimes spontaneously switch between them (for reviews, see, e.g., Dember & Fowler, 1958; Richman, Dember, & Kim, 1986). This phenomenon may also be tied to sequential dependencies in many natural foraging environments (Erev & Haruvy, 2016). In other words, choice diversification strategies—such as probability matching or spontaneous alternation—can be regarded as exploratory strategies that are adaptive in many real-world settings entailing some form of dynamic change (e.g., dependencies of rewards on prior choices as considered here or environmental changes that are independent of decision makers' actions).

A conclusion regarding whether organisms generally tend to over- or under-explore their environment cannot, however, be drawn from the literature. On the one hand, research on explore–exploit problems has suggested that an initial "bias" toward exploratory choice may result from the higher costs associated with assuming consistency when the environment is in fact dynamic, relative to the costs of occasional exploration in an unchanging context (Navarro & Newell, 2014; Navarro, Newell, & Schulze, 2016). Yet on the other hand, in related dynamic sequential choice contexts, people often appear to not explore enough but to persist with strategies that are suboptimal (e.g., Frey, Rieskamp, & Hertwig, 2015; Lejuez et al., 2002). This opposite tendency to "under-explore" may be related to an asymmetry in the impact of good versus bad experiences on choice—the "hot stove" effect (Denrell & March, 2001). That

is, good outcomes increase the likelihood of repeating a choice whereas bad ones lead to avoidance, which can distort a decision maker's knowledge about which line of action yields the best outcomes. When people are provided with this missing information—e.g., full feedback about forgone or future outcomes—they have been found to explore more adequately (Frey et al., 2015).

To conclude, we have empirically examined adaptive human choice diversification and persistence in the presence/absence of choice–outcome dependencies that were introduced by the retention/withdrawal of reward across trials. Our findings demonstrate that, given sufficient feedback, some (but not all) people learn to choose effectively in both contexts. This research contributes to a broadened understanding of environmental determinants for adaptive decision making by bridging theoretical concepts of human probability learning and animal operant conditioning. An interesting avenue for future research would be to explore the relationship between animal and human behavior in equivalent tasks, thus furthering the establishment of theoretical connections between research on probability learning and the matching law in concurrent reinforcement.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110. http://dx.doi.org/10.1016/0749-5978(86)90046-4

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369–1378. http://dx.doi.org/10.3758/BF03206327

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, *22*(1), 231–242. http://dx.doi.org/10.1901/jeab.1974.22-231

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, *32*(2), 269–281. http://dx.doi.org/10.1901/jeab.1979.32-269

Borrero, J. C., Crisolo, S. S., Tu, Q., Rieland, W. A., Ross, N. A., Francisco, M. T., & Yamamoto, K. Y. (2007). An application of the matching law to social dynamics. *Journal of Applied Behavior Analysis*, *40*(4), 589–601. http://dx.doi.org/10.1901/jaba.2007.589-601

Borrero, J. C., & Vollmer, T. R. (2002). An application of the matching law to severe problem behavior. *Journal of Applied Behavior Analysis*, *35*(1), 13–27. http://dx.doi.org/10.1901/jaba.2002.35-13

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. New York, NY: Springer-Verlag.

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*(3), 253–262. http://dx.doi.org/10.1037/1040-3590.14.3.253

Dember, W. N., & Fowler, H. (1958). Spontaneous alternation behavior. *Psychological Bulletin*, *55*(6), 412–428. http://dx.doi.org/10.1037/h0045446

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538. http://dx.doi.org/10.1287/orsc.12.5.523.10092

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

Erev, I., & Haruvy, E. (2016). Learning and the economics of small decisions. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (2nd ed., pp. 638–716). Princeton, NJ: Princeton University Press.

Estes, W. K. (1964). Probability learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 89–128). New York, NY: Academic Press.

Fam, J., Westbrook, F., & Arabzadeh, E. (2015). Dynamics of pre- and post-choice behaviour: Rats approximate optimal strategy in a discrete-trial decision task. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1803), 20142963.

Frey, R., Rieskamp, J., & Hertwig, R. (2015). Sell in May and go away? Learning and risk taking in nonmonotonic decision problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 193–208. http://dx.doi.org/10.1037/a0038118

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422. http://dx.doi.org/10.1016/j.cognition.2008.09.007

Gal, I., & Baron, J. (1996). Understanding repeated simple choices. *Thinking & Reasoning*, *2*(1), 81–98. http://dx.doi.org/10.1080/135467896394573

Goodnow, J. J. (1955). Determinants of choice-distribution in two-choice situations. *The American Journal of Psychology*, *68*(1), 106–116. http://dx.doi.org/10.2307/1418393

Green, C. S., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the National Academy of Sciences*, *107*(37), 16401–16406. http://dx.doi.org/10.1073/pnas.1001709107

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. http://dx.doi.org/10.1007/BF02289823

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293–313. http://dx.doi.org/10.1016/j.cognition.2009.03.013

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267–272. http://dx.doi.org/10.1901/jeab.1961.4-267

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243–266. http://dx.doi.org/10.1901/jeab.1970.13-243

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior*, *24*(1), 107–116. http://dx.doi.org/10.1901/jeab.1975.24-107

Houston, A. I., & McNamara, J. (1981). How to maximize reward rate on two variable-interval paradigms. *Journal of the Experimental Analysis of Behavior*, *35*(3), 367–396. http://dx.doi.org/10.1901/jeab.1981.35-367

James, G., & Koehler, D. J. (2011). Banking on a bad bet: Probability matching in risky choice is linked to expectation generation. *Psychological Science*, *22*(6), 707–711. http://dx.doi.org/10.1177/0956797611407933

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (2016). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*. Advance online publication. http://dx.doi.org/10.3758/s13428-016-0739-8

Jensen, G., & Neuringer, A. (2008). Choice as a function of reinforcer "hold": From probability learning to concurrent reinforcement. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*(4), 437–460. http://dx.doi.org/10.1037/0097-7403.34.4.437

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In IEEE, Neural Networks Council (Ed.), *IEEE International Conference on Neural Networks Proceedings* (pp. 1942–1948). Perth, Australia: IEEE.

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667–676. http://dx.doi.org/10.3758/MC.38.6.667

Koehler, D. J., & James, G. (2014). Probability matching, fast and slow. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 103–131). San Diego, CA: Elsevier.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. http://dx.doi.org/10.1901/jeab.2005.110-04

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*(2), 75–84. http://dx.doi.org/10.1037/1076-898x.8.2.75

MacDonall, J. S. (1988). Concurrent variable-ratio schedules: Implications for the generalized matching law. *Journal of the Experimental Analysis of Behavior*, *50*(1), 55–64. http://dx.doi.org/10.1901/jeab.1988.50-55

MacGregor, R. (1999). Fortune smiled upon us. In S. Dryden (Ed.), *Total Gretzky: The magic, the legend, the numbers* (pp. 14–36). Toronto, Canada: McClelland & Stewart.

McDowell, J. J. (2013). On the theoretical and empirical status of the matching law and matching theory. *Psychological Bulletin*, *139*(5), 1000–1028. http://dx.doi.org/10.1037/a0029924

Morey, R. D., & Rouder, J. N. (2015). BayesFactor (Version 0.9.12-2) [Computer software]. Retrieved from http://bayesfactorpcl.r-forge.r-project.org

Navarro, D. J., & Newell, B. R. (2014). Information versus reward in a changing world. In P. Bellow, M. Guarani, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1054–1059). Austin, TX: Cognitive Science Society.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. http://dx.doi.org/10.1016/j.cogpsych.2016.01.001

Newell, B. R., Koehler, D. J., James, G., Rakow, T., & van Ravenzwaaij, D. (2013). Probability matching in risky choice: The interplay of feedback and strategy availability. *Memory & Cognition*, *41*(3), 329–338. http://dx.doi.org/10.3758/s13421-012-0268-3

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*(6), 1133–1139. http://dx.doi.org/10.3758/BF03193102

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*(6432), 56–58. http://dx.doi.org/10.1038/364056a0

Otto, A. R., & Love, B. C. (2010). You don't want to know what you're missing: When information about forgone rewards impedes dynamic decision making. *Judgment and Decision Making*, *5*(1), 1–10.

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*(2), 274–279. http://dx.doi.org/10.1016/j.cognition.2010.11.009

Peterson, C. R., & Ulehla, Z. J. (1965). Sequential patterns and maximizing. *Journal of Experimental Psychology*, *69*(1), 1–4. http://dx.doi.org/10.1037/h0021597

Rakow, T., Newell, B. R., & Zougkou, K. (2010). The role of working memory in information acquisition and decision making: Lessons from the binary prediction task. *The Quarterly Journal of Experimental Psychology*, *63*(7), 1335–1360. http://dx.doi.org/10.1080/17470210903357945

Reed, D. D., Critchfield, T. S., & Martens, B. K. (2006). The generalized matching law in elite sport competition: Football play calling as operant choice. *Journal of Applied Behavior Analysis*, *39*(3), 281–297. http://dx.doi.org/10.1901/jaba.2006.146-05

Richman, C. L., Dember, W. N., & Kim, P. (1986). Spontaneous alternation behavior in animals: A review. *Current Psychology*, *5*(4), 358–391. http://dx.doi.org/10.1007/BF02686603

Rothstein, J. B., Jensen, G., & Neuringer, A. (2008). Human choice among five alternatives when reinforcers decay. *Behavioural Processes*, *78*(2), 231–239. http://dx.doi.org/10.1016/j.beproc.2008.02.016

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *Journal of Neuroscience*, *29*(48), 15104–15114. http://dx.doi.org/10.1523/JNEUROSCI.3524-09.2009

Schulze, C., & Newell, B. R. (2016). More heads choose better than one: Group decision making can eliminate probability matching. *Psychonomic Bulletin & Review*, *23*(3), 907–914. http://dx.doi.org/10.3758/s13423-015-0949-6

Schulze, C., van Ravenzwaaij, D., & Newell, B. R. (2015). Of matchers and maximizers: How competition shapes choice under risk and uncertainty. *Cognitive Psychology*, *78*, 78–98. http://dx.doi.org/10.1016/j.cogpsych.2015.03.002

Serences, J. T. (2008). Value-based modulations in human visual cortex. *Neuron*, *60*(6), 1169–1181. http://dx.doi.org/10.1016/j.neuron.2008.10.051

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250. http://dx.doi.org/10.1002/bdm.413

Sims, C. R., Neth, H., Jacobs, R. A., & Gray, W. D. (2013). Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review*, *120*(1), 139–154. http://dx.doi.org/10.1037/a0030850

Staddon, J. E. R., Hinson, J. M., & Kram, R. (1981). Optimal choice. *Journal of the Experimental Analysis of Behavior*, *35*(3), 397–412. http://dx.doi.org/10.1901/jeab.1981.35-397

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

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118. http://dx.doi.org/10.1111/1467-6419.00106

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. http://dx.doi.org/10.3758/BF03206482

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243–251. http://dx.doi.org/10.3758/BF03194383

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1640–1658. http://dx.doi.org/10.1037/a0028146

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

Yechiam, E., & Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. *Journal of Mathematical Psychology*, *51*(2), 75–84. http://dx.doi.org/10.1016/j.jmp.2006.11.002

Yechiam, E., Rakow, T., & Newell, B. R. (2015). Super-underweighting of rare events with repeated descriptive summaries. *Journal of Behavioral Decision Making*, *28*(1), 67–75. http://dx.doi.org/10.1002/bdm.1829

Appendix

Model Specifications

Baseline. All strategy models were compared with a baseline model that fitted a constant and statistically independent choice probability of selecting the initially more probable option, $p(H)$, $p(L) = 1 - p(H)$, to each participant's choices across trials (see, e.g., Busemeyer & Stout, 2002). The only free parameter in this model was individually estimated from participants' actual choice proportions in our experiment; it provides no conceptual account of the decision processes involved.

Win-stay lose-shift. We applied a probabilistic modification of the deterministic WSLS rule by allowing for free variation of the probability of repeating the same choice following a success, denoted as $p_{t+1}(stay|win_t)$, and the probability of changing to the other option following a failure, denoted as $p_{t+1}(shift|lose_t)$.

Reinforcement learning. We considered a simple learning model that assumes a decision maker to gradually establish propensities toward choice options, denoted as $q_t(i)$ for option *i* on trial *t*, based on the discrepancy between expected and actual monetary rewards (0 or 1 cent in our task; see Method) on a given choice trial:

$$q_t(i) = q_{t-1}(i) + \alpha \cdot [r_t(i) - q_{t-1}(i)]. \quad (3)$$

Here, propensities for options for which outcome feedback is available are updated in increments of the learning rate α times the prediction error in brackets. The precision with which the preferred option is then selected on the following trial is determined on the basis of a softmax choice rule (Sutton & Barto, 1998) contingent on the sensitivity parameter θ:

$$p_{t+1}(i) = \frac{e^{\theta \cdot q_t(i)}}{e^{\theta \cdot q_t(j)} + e^{\theta \cdot q_t(i)}}, \ \theta = 3^{10 \cdot c} - 1. \quad (4)$$

An exponential transformation of θ was employed to allow variation of choice precision between random guessing (for $\theta \approx 0$) and strictly deterministic selection of the preferred alternative (for $\theta > 700$) within narrow bounds of the sensitivity constant *c* (see, e.g., Yechiam & Ert, 2007).

Choice pattern. We introduced a strategy model encompassing a range of regular choice rules that vary in the extent to which they are deterministic. This model assumes that each decision maker has a preferred successive choice run length for each option, denoted as β_i for option *i*, after which the decision maker becomes indifferent between the options. These preferred choice run lengths define the regularity of the implemented choice sequence. On each trial *t*, the selected option's preferred run length parameter is evaluated against the number of previous choices of that option, resulting in a deviation measure for the currently exploited option given by

$$d_t(i) = prev_t(i) - \beta_i. \quad (5)$$

This deviation determines the probability with which the decision maker shifts to the unselected option on the next trial by defining the inflection point of a logistic choice function, such that

$$p_{t+1}(shift) = \frac{1}{1+e^{-\theta \cdot d_t(i)}}, \quad \theta = 3^{10 \cdot c} - 1. \quad (6)$$

According to this choice rule, positive/negative deviations from preferred successive choice length result in a greater/smaller than chance probability that a decision maker will shift away from the previously selected option. This relationship between choice probability and current choice run length is illustrated in Figure A1 for different values of the preferred choice run length parameter. Additionally, the sensitivity parameter, denoted as θ, describes choice precision in terms of adherence to the specified choice rule. That is, the choice rule precision parameter governs the slopes of the sigmoid choice probability function as illustrated in Figure A1. As for the reinforcement learning model, we used an exponential transformation of θ to allow random to deterministic variation of choice pattern adherence within narrow bounds of the sensitivity constant *c*. Thus, this choice pattern model has three parameters—preferred choice run lengths for each option after which preference to both options becomes

indifferent (β_{high} and β_{low}) and overall choice rule precision (*c*)—and describes choice probabilities solely on the basis of prior choice regularities, irrespective of obtained rewards.

[insert Figure A1 here]

Parameter Estimation and Model Evaluation

We estimated parameter values that maximized the summed log-likelihood of each model separately for each participant on the basis of its accuracy in predicting the observed choice sequence one trial ahead. Optimization was carried out with an iterative particle swarm optimization method (Kennedy & Eberhart, 1995) as described in detail in Schulze et al. (2015). Parameter estimation was constrained by the following parameter bounds: for the baseline model, the constant probability of choosing the initially higher option, $p(H)$, was constrained between 0 and 1; for the WSLS model, $p(stay|win)$ and $p(shift|lose)$ were allowed to vary freely between 0 and 1; the reinforcement learning parameter α and the sensitivity constant *c* were constrained between 0 and 1; and finally, for the choice pattern model, the preferred cycle length parameters β_i could vary between 0 and 500 (i.e., the maximum preferred cycle length for a choice task with a total of 500 trials) and the sensitivity constant *c* was again constrained between 0 and 1. To capture which strategy model best described participants' choices, we categorized each participant based on the relative strength of evidence in favor of each strategy model in terms of AIC model weights (*wAIC*). These model weights are interpretable as the conditional probability that a strategy model is the best in a set of candidate models, given the data of a particular participant and the candidate models (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004). Accordingly, strategy use was categorized by identifying the highest *wAIC* for each participant. This measure discriminated between strategy model fits for all participants.

Tables

Table 1

Average reward proportion expected for choice sequences consisting of y (row number) consecutive choices of the high and x (column number) consecutive choices of the low option

Consecutive choices of the high option \ Consecutive choices of the low option	**0**	**1**	**2**	**3**	**4**	**…**
0		.30	.30	.30	.30	
1	.70	.710	.594	.525	.482	
2	.70	.756	.658	.590	.542	
3	.70	.767	.687	.625	.580	
4	.70	**.768**	.701	.646	.604	
5	.70	.765	.708	.659	.620	
6	.70	.761	.711	.668	.632	
7	.70	.757	.713	.673	.640	⋮
8	.70	.752	.713	.677	.646	
9	.70	.748	.713	.680	.651	
10	.70	.745	.713	.682	.656	
11	.70	.741	.712	.684	.659	
12	.70	.738	.712	.685	.662	
13	.70	.736	.711	.687	.664	
14	.70	.734	.711	.687	.666	
15	.70	.732	.710	.688	.668	
…	…					

Note. Dark shading signifies diversification sequences that yield higher reward rates than the average reward probability expected from static probability maximizing, $p(\mathrm{H}) = .70$. The maximum average reward is attained by adopting a choice sequence in which four consecutive high choices are followed by one low choice (marked in black).

Table 2

Number of participants classified as using effective diversification, probability maximizing, or other strategies during the final block of 100 trials, and association between strategy classification and reward hold manipulation within and across feedback conditions

		Classified strategy			Statistics		
Feedback type	Environment	Effective diversification	Probability maximizing	Other strategies	χ^2	*p*	*BF*
Partial feedback	Hold	5	1	17	3.98	.137	0.70
	No hold	2	5	16			
Full feedback	Hold	10	1	12	**13.80**	**.001**	**180.40**
	No hold	1	9	13			
Across feedback conditions	Hold	15	2	29	**17.00**	**< .001**	**479.32**
	No hold	3	14	29			

Note. For all chi-square tests $df = 2$. *BF*s were calculated under the assumption that rows were sampled as independent multinomials with their total fixed. Significant statistics are marked in bold.

Table 3

Mean (SD) parameter estimates for each strategy model and experimental condition.

	Hold		No hold	
	Partial feedback	Full feedback	Partial feedback	Full feedback
WSLS				
p(stay|win)	.72 (.15)	.76 (.13)	.79 (.18)	.87 (.10)
p(switch|lose)	.65 (.22)	.57 (.19)	.37 (.18)	.29 (.20)
Reinforcement learning				
Learning rate (α)	.49 (.40)	.36 (.39)	.51 (.42)	.10 (.23)
Choice sensitivity (c)	.12 (.03)	.23 (.26)	.21 (.23)	.28 (.19)
Choice pattern				
Run length high (β_{high})	268.98 (247.55)	191.96 (231.89)	349.57 (221.83)	413.73 (192.25)
Run length low (β_{low})	11.72 (36.78)	8.97 (32.54)	155.08 (184.30)	101.25 (146.60)
Choice sensitivity (c)	.008 (.013)	.007 (.007)	.002 (.004)	.003 (.008)

Figures

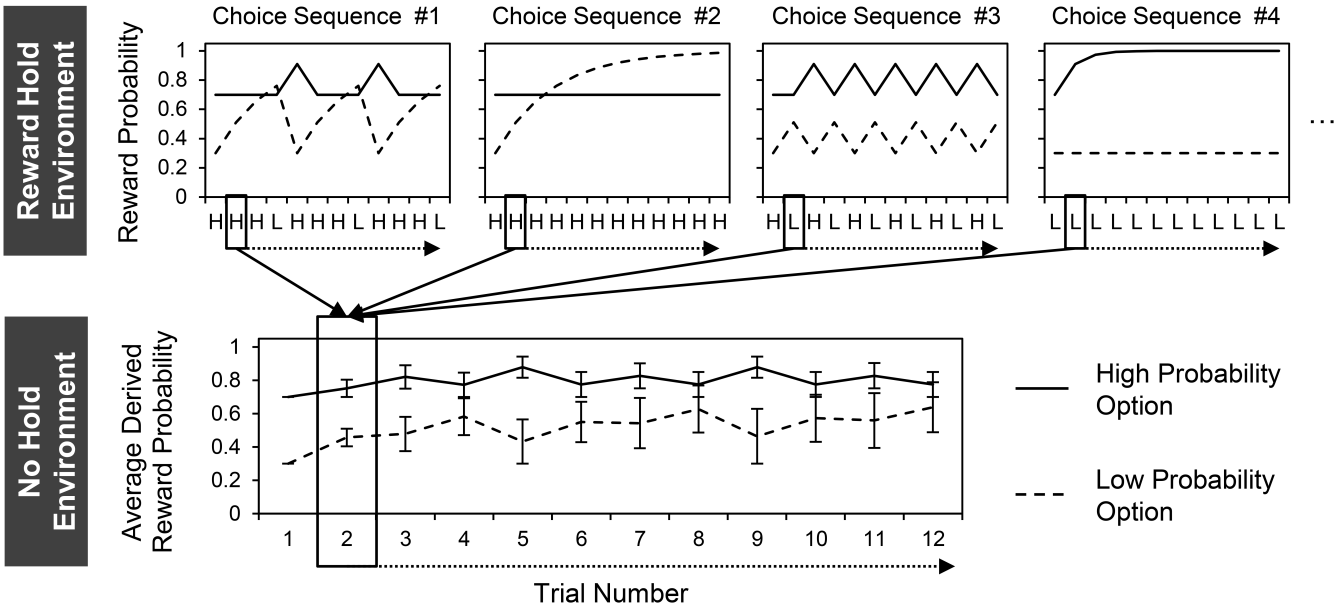

Figure 1. Overview of the task manipulation (reward hold vs. no hold) and ensuing reward structure. The top part shows reward probability changes as a function of four hypothetical 12-trial choice sequences (letters H/L denote choices of the option with the initially higher/lower outcome probability) under reward hold. The bottom part shows reward probability changes derived from these hypothetical choices by averaging across multiple sequences and their associated outcome probabilities for both options under no hold. This procedure renders the reward probabilities comparable between environments, but these probabilities either change in response to current choices (top part, reward hold) or irrespective of current choices (bottom part, no hold). See the Method section for details of the characteristics and construction of both reward structures; the reward probabilities displayed here are purely illustrative.

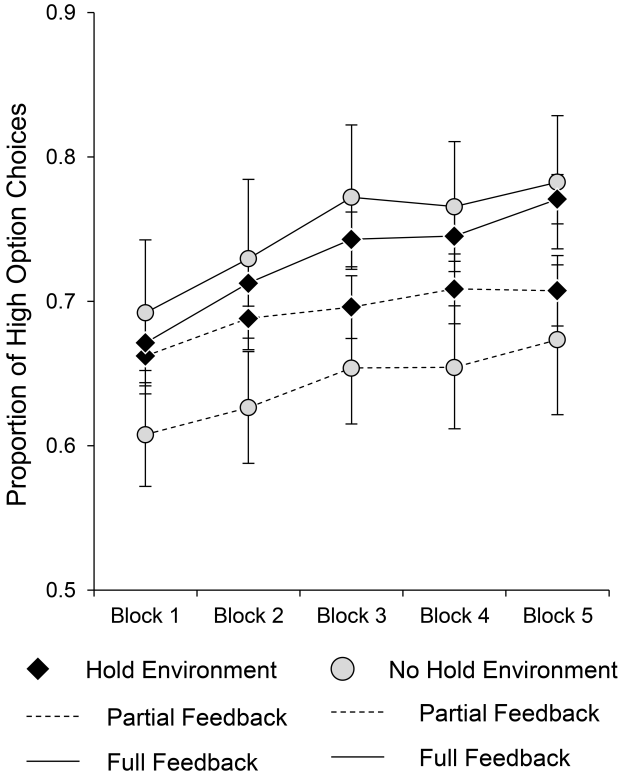

Figure 2. Mean (± standard error) proportion of participants' choices of the high-probability option in each block of 100 trials by reward hold and feedback condition.

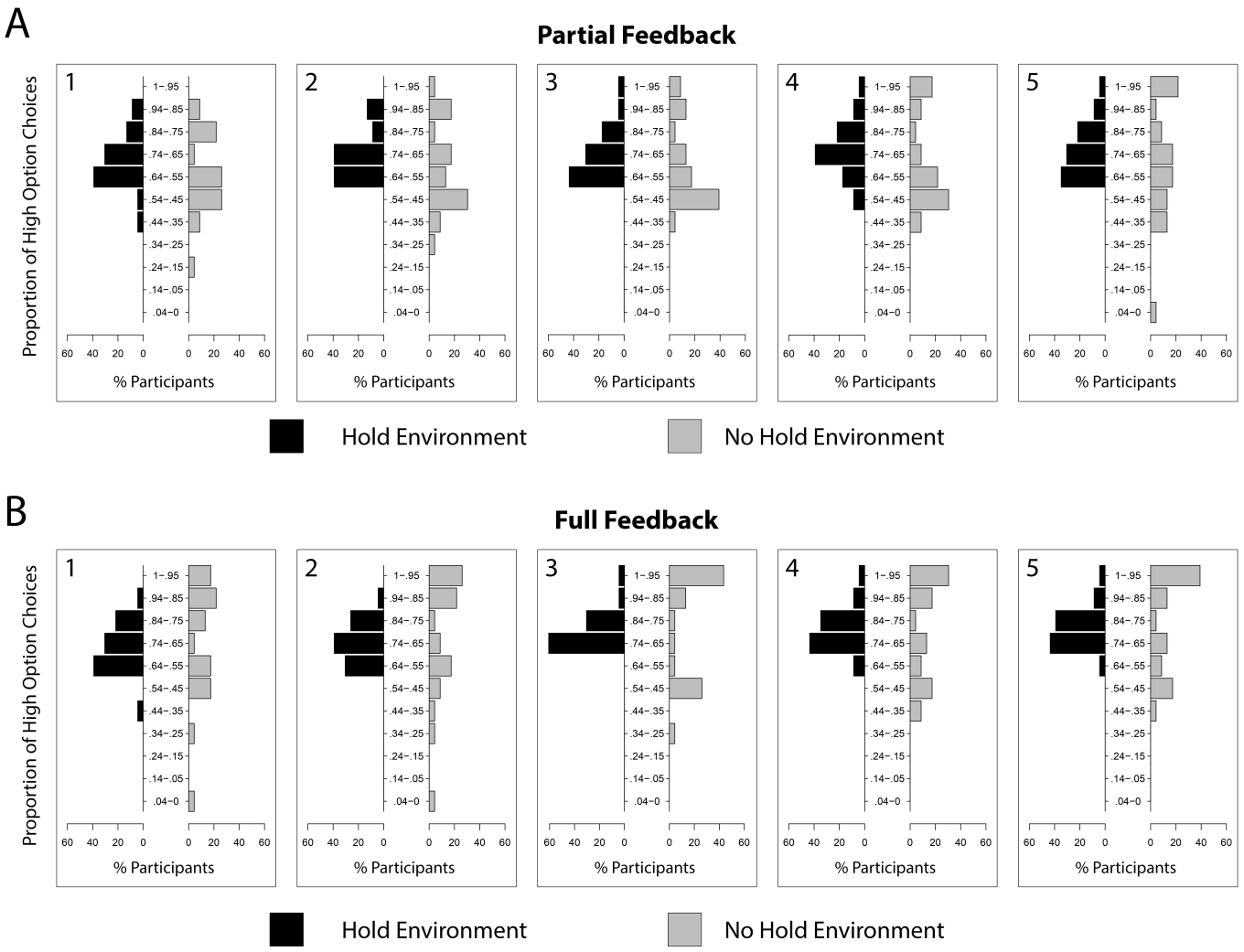

Figure 3. Distributions of individual participants' proportion of high-probability option choices for the partial feedback conditions (A) and the full feedback conditions (B). Each graph displays the distributions of participants' high-option choices during one block of 100 trials, as indicated by the number in each left upper corner. The black bars depict the distributions of participants under reward hold; the gray bars, under no hold. The top category represents probability maximizing; the third category from the top approximately represents effective diversification.

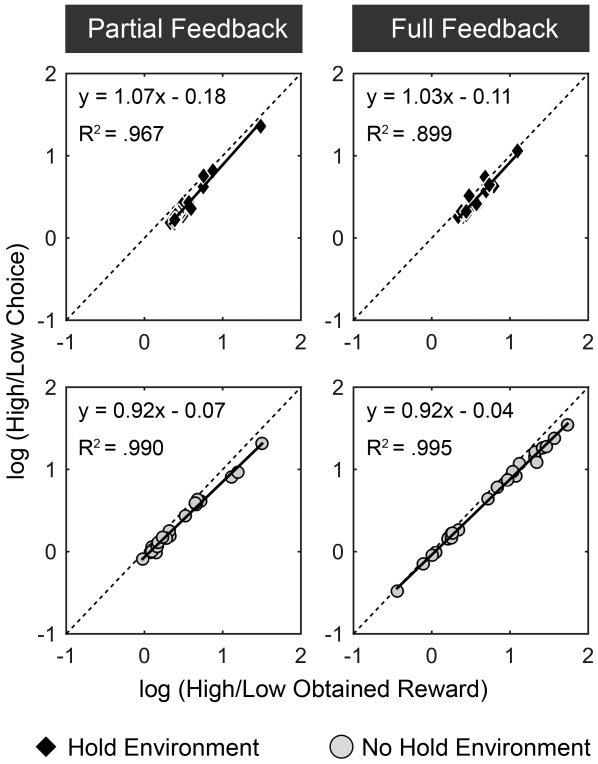

Figure 4. Log response ratios as a function of log obtained reward ratios. Each black diamond/gray circle represents one participant in the reward hold/no hold environment, respectively; the two feedback conditions are plotted separately. Solid lines represent least squares linear regression fits of the logarithmic form of the generalized matching law (Equation 1) to participants' choice data in each experimental condition. Dashed lines represent perfect matching.

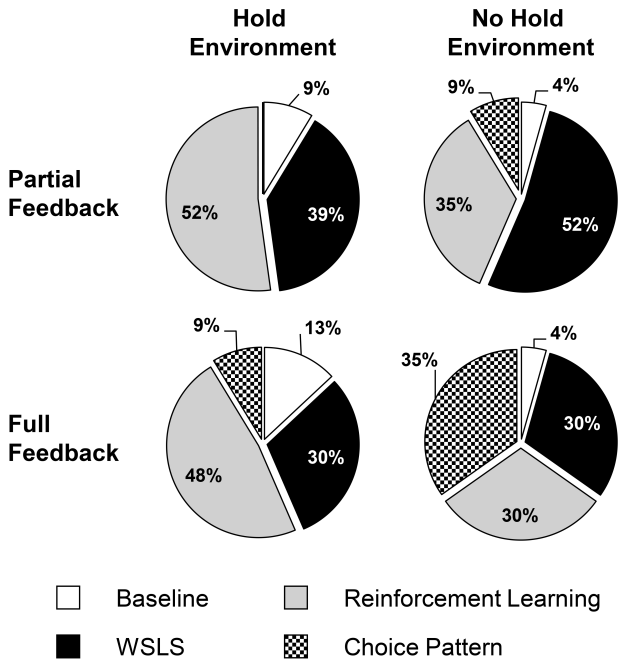

Figure 5. Proportion of participants in each experimental condition for whom the specified model provided the best fit to the data. Model classifications were based on AIC model weights.

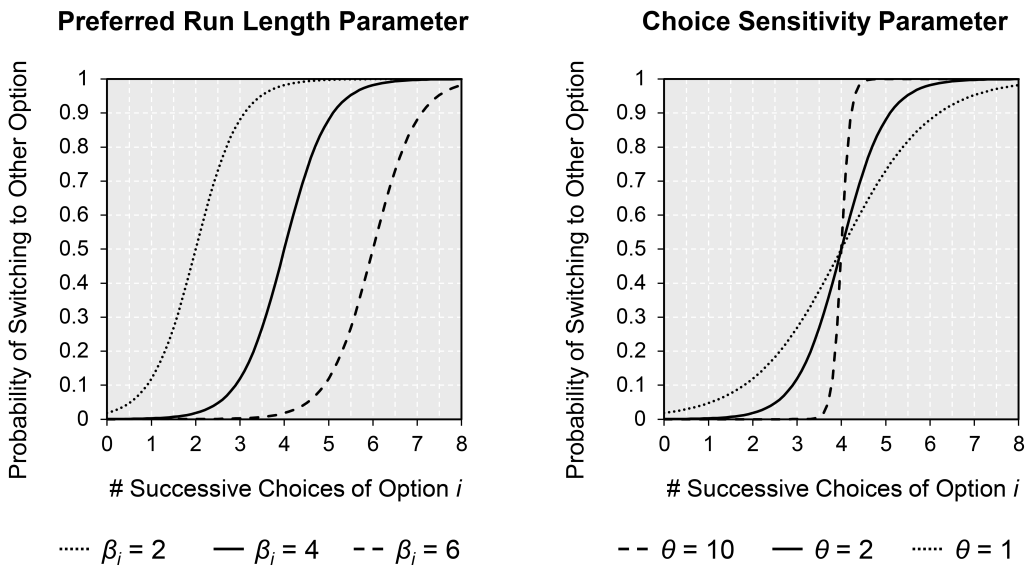

Figure A1. Probability of switching choice options as a function of choice run length for different values of the preferred run length model parameter β_i (left panel) and the choice rule precision parameter θ (right panel). The preferred choice run length parameter β_i determines the inflection point of the choice probability function (all curves in the left panel are shown with $\theta = 2$); the choice precision parameter θ defines the slope of the choice probability function (all curves in the right panel are shown with $\beta_i = 4$).