Match Me if You Can:
How Smart Choices are Fueled by Competition

Christin Schulze (c.schulze@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW 2052 Australia

Don van Ravenzwaaij (d.vanravenzwaaij@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW 2052 Australia

Ben R. Newell (ben.newell@ unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW 2052 Australia

Abstract

In a world of limited resources, scarcity and rivalry are central challenges for decision makers. We examine choice behavior in competitive probability learning environments that reinforce one of two strategies. The optimality of a strategy is dependent on the behavior of a computerized opponent: if the opponent mimics participant choices, probability *matching* is optimal; if the opponent is indifferent, probability *maximizing* is optimal. We observed accurate asymptotic strategy use in both conditions suggesting participants were sensitive to the differences in opponent behavior. Moreover, the results emphasize that 'irrational' probability matching can be adaptive once such competitive pressures are taken into account. The application of reinforcement learning models to the data suggests that computational conceptualizations of opponent behavior are critical to account for the observed divergence in strategy adoption.

Keywords: Decision making; Probability matching; Reinforcement learning; Evolutionary psychology; Mathematical modeling

Introduction

Competition is a pervasive characteristic of the world – plants compete for light, water and pollination; animals are in continual competition for territory, food and mating; and even as humans we are constantly competing in sports, for social standing and companionship. Considering the ubiquity of competitive pressures in virtually all aspects of our lives, their crucial impact on the development of adaptive decision strategies in a broad range of contexts may seem self-evident. And yet, prior research has concentrated on assessing the rationality of numerous choice phenomena primarily by focusing on individual decision makers in social isolation. Consequently, observed choice inconsistencies are frequently dismissed as suboptimal with little or no regard for their adaptive potential in ecologically valid settings (see e.g., Todd & Gigerenzer, 2007).

One such extensively studied choice anomaly is the tendency to proportionately match choices to outcome probabilities in repeated binary decisions, a phenomenon known as probability matching (for a review see Vulkan, 2000). In a typical setup a decision maker repeatedly has two choice options available, one of which is the correct choice with greater probability than its alternative, e.g. $p(\mathrm{A}_1) = .7$ and $p(\mathrm{A}_2) = .3$, and correct predictions are rewarded with monetary payoffs. Assuming the outcome probabilities are stationary and irrespective of prior events or subjects' behavior, A_1 is the superior choice option throughout and, following an initial period of probability learning, should be chosen exclusively. By contrast, the frequently observed probability (over-) matching tendency results in inferior prediction accuracies and payoffs and is therefore considered fallacious within context-independent interpretations of rational choice behavior (Vulkan, 2000).

Probability Matching in Competitive Environments

What seems irrational in individualized context-free environments, however, can be optimal in ecologically more valid situations that take prospective social interactions into account (Gallistel, 1990; Gigerenzer, 2000). That is to say, when decision makers seek to exploit limited resources under *natural circumstances* (e.g. forage for food or make money), they are rarely alone but typically in fierce competition for the exploitation of these resources with other agents. The more individual agents then choose the seemingly richest resource, the smaller each one's share. In nature, this situation cannot remain stable as natural selection would favor those agents who sometimes chose options with potentially scarce resources that are exploitable under less competition (Gallistel, 1990).

Following this line of argument, it has been suggested that agents should distribute their choices among resources relative to their reward potentials, i.e. adopt a probability matching strategy, to create an equilibrated evolutionary stable situation that does not give rise to conditions selecting against it (Gallistel, 1990). Evidence for such behavior has been provided by experiments that studied groups of animals in the wild, e.g. foraging behavior of ducks on a lake (Harper, 1982) and fish in a tank (Godin & Keenleyside, 1984).

Our aim was to examine the role of competitive pressures for the facilitation of optimal decision making in simple binary human choice contexts. Specifically, we wanted to assess potential benefits of probability matching under the premise that competitive conditions reinforce its superiority. Following the logic of natural foraging situations, we designed a choice environment in which each decision maker competes against a computerized opponent for the exploitation of a monetary resource that an indifferent 'nature' has placed at one of two choice options with static probabilities. When the two agents converge on the same choice, potential rewards are split evenly between them.

In this competitive context, the success of any strategy largely depends on the behavior of the opponent. Under the assumption that the competitor is attentive towards the decision maker's choice behavior and imitates her course of action, probability matching is an optimal strategy. The prevalence of aggregative behavior in a broad range of natural group settings, e.g. in flocking behavior of birds, shoaling of fish, swarming of insects and herd behavior of land animals (Allee, 1978), suggests that a strategy-mirroring opponent creates competitive conditions closely in line with real life ecological pressures. Thus, in one experimental condition each opponent's choice probabilities are close imitations of participant behavior which renders probability matching optimal (see appendix). In a second condition (between-subjects), each participant is paired with a computer opponent who is indifferent towards subjects' choices thereby making exclusive preference for the more profitable resource (i.e. probability maximizing) the optimal strategy. This is the case because sporadic choices by the participant to the lesser option will not tempt this indifferent opponent to replicate deviant behavior but will merely result in relinquished earning potential for the participant.

By manipulating opponent behavior as described, we created two competitive choice environments that differed solely in the extent to which participants had influence on their competitors' behavior. Thus, we can assess the role of the qualitative nature of competition for the facilitation of adaptive decision making. Given the availability of sufficient feedback (Shanks, Tunney, & McCarthy, 2002), we predicted that choices will converge on the respective optimal strategy in both environments as learning progresses, i.e. probability matching when competing against a mimicking opponent and probability maximizing when encountering an indifferent opponent.

Models of Learning under Competition

To shed more light on the nature of underlying learning processes within such competitive environments, we discuss the applicability of reinforcement learning models proposed for similar choice settings, e.g. learning in experimental games (Erev & Roth, 1998), learning in the Iowa Gambling task (Yechiam & Busemeyer, 2005) and strategy selection learning (Rieskamp & Otto, 2006), to our experiments and outline potential adaptions of such models to account for the competitive pressures examined here. Such models typically include assumptions regarding three main components (see e.g., Sutton, 1998): a utility function that specifies the goal of the learning problem; a learning rule which establishes propensities for each choice option; and a choice rule defining the course of action given current propensities. Here we examine two learning models postulating different conceptualizations of the utility formation process.

Utility Function In a learning environment where an agent's primary goal is maximization of total payoffs, the utility of a choice is typically directly specified by its associated monetary reward (e.g. Rieskamp & Otto, 2006):

$$u_t(i) = r_t(i), \tag{1}$$

where $u_t(i)$ corresponds to the utility of the monetary gains $r_t(i)$ associated with choice *i* on trial *t*, namely, in our task, 0, 2 or 4 cents for no, split and full payoffs (see below). The focus on monetary gains for the evaluation of choice utilities has left systematic investigations of a wider range of factors potentially influencing this important model component largely unexplored (with few notable exceptions, e.g. Janssen & Gray, 2012; Singh, Lewis, & Barto, 2009). This is the case even though various additional motivational sources of utility are conceivable: e.g. avoidance of boredom associated with repetitive tasks (Keren & Wagenaar, 1985) or task completion time (Gray, Sims, Fu, & Schoelles, 2006). Relating this prevalent negligence to the competitive task employed here, we argue that describing utilities in terms of monetary rewards only confounds two discrete learning goals that drive learning in this context, namely, correctly assessing the profitability of an option *and* attending to the opponent's choices. In fact, monetary based utilities understate the crucial role differential causes of opponent behavior play when subjects face an imitative or an indifferent competitor. That is to say, different opponent strategies necessitate divergent learning goals: if an opponent is identified as attentive, deciding on a course of action requires consideration of ways to influence and outsmart that other agent; if, on the other hand, the competitor is indifferent, the impact of opposing actions on one's own decisions should be strongly discounted.

Incorporating these aspects into the learning model we propose a utility function that disentangles the two learning goals present in our task and allows direct estimation of the importance decision makers attribute to the choice strategies they observe in their competitors compared to the importance they ascribe to choosing the better option:

$$u_t(i) = [\beta \cdot g_t(i)] + [(1 - \beta) \cdot s_t(i)]. \tag{2}$$

Here, the utility $u_t(i)$ of a choice, is expressed as the weighted sum of its accuracy $g_t(i)$ (0 for incorrect and 1 for correct guesses) and the choice of the competitor $s_t(i)$ (-1 for converging choices and 1 for incongruent choices) on any given trial. The additional free parameter β determines the weight a subject assigns to choosing the correct option as compared to outsmarting their competitor in terms of choosing the opposite line of action. For $\beta = 1$ subjects value the accuracy of their choices only, whereas for $\beta = .5$ the importance of correct choices and outwitting the competitor are weighted equally. We predict that balancing these two requirements of the task would be more pronounced when facing an imitative competitor, thus $M_{\beta,indifferent} > M_{\beta,mimicry}$, and that learning models considering these differential challenges of the task would account for the data more thoroughly.

Updating and Choice Rule Adjustment of propensities follows a delta learning rule commonly employed in similar decision tasks (e.g., Yechiam & Busemeyer, 2005):

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

Here, initial propensities towards both options are assumed to equal zero and are then gradually updated in increments of the learning rate α based on the prediction error in brackets. As outcomes are mutually exclusive in our task, propensities for both options are updated simultaneously regardless of the actual choice on any given trial. An agent's probability of choosing either option is determined by these propensities following an exponential 'softmax' choice rule:

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

where the sensitivity parameter θ governs the precision with which the preferred option is chosen. If $\theta = 0$, decisions are made at random, i.e. $p_t(i) = p_t(j) = .5$, whereas large sensitivity parameter values ($\theta \to \infty$) correspond to strictly deterministic choices to the option with the higher propensity. Following Yechiam & Ert (2007), an exponential transformation of θ was employed to allow variation of choice sensitivities between random guessing (for $\theta \approx 0$) to fully deterministic (for $\theta > 700$) within narrow bounds of c, which denotes the sensitivity constant constrained between 0 and 1.

Method

Participants

Fifty (35 female) undergraduate students from the University of New South Wales (mean age 18.9, *SD* = 1.2 years) participated in this experiment in return for course credit and performance based monetary compensation.

Decision Task

A standard probability learning paradigm involving repeated binary decisions with mutually exclusive outcomes over 500 choice trials was employed. Choice alternatives were represented by two light bulbs displayed on a computer screen and programmed to illuminate with probabilities of .7 and .3, counterbalanced across participants for left and right choice options. Correct predictions were rewarded with 4 cents (1 AUD = .95 USD). Choices were made while competing against a computerized opponent and when both agents converged on the correct response, the payoff was evenly split between them, i.e. each agent received 2c.

Design

We employed a 2 x 5 mixed model design with opponent type (mimicry or indifferent) as between-subjects factor and trial block (five blocks of 100 trials each) as within-subjects factor. The dependent measure was the proportion of choices to the more profitable choice option. For the mimicry group, the choice sequence of each opponent was computed one step ahead by equating the opponent's choice probabilities on each trial with the choice proportions the subject had displayed during the past ten trials. For example, when a participant chose the more profitable option on 7 out of the past 10 trials, her opponent's probability of choosing the same option on the subsequent trial was .7.[1] This algorithm creates opponent behavior that probabilistically mimics participants' choices.

By contrast, the choice sequence of each opponent for subjects in the indifferent condition was computed irrespective of participants' choices. Instead, each subject played against an opponent whose set of choice *probabilities* simply repeated those of an opponent encountered by another subject in the mimicry condition.

Procedure

Subjects were asked to predict which of two light bulbs would illuminate over a series of trials while attempting to earn as much money as possible. Instructions indicated that the lighting sequence was random, i.e. no pattern or system existed which made it possible to correctly predict the outcome throughout, and that the outcome probabilities of both choice options remained constant during the entire experiment. Additionally, participants in both conditions were informed that a computerized opponent with learning abilities such as their own and no initial information about the lighting frequencies was monitoring their choices and adapting to their skill level. On each trial, predictions were made simultaneously by both participant and opponent and followed by feedback about the other agent's choice and the outcome, i.e. one light bulb lit up.

Upon completion of every block of 100 trials a self-paced pause interrupted the experiment during which block feedback was provided and a short message reminded participants that the lighting sequence was random. Subjects were told: "In this game you earned X$. Using an optimal strategy you could have earned at least Y$.", where X represented the actual earnings of that block and Y was computed by an optimizing algorithm (Shanks et al., 2002). This algorithm was set to probability matching in the mimicry opponent condition and probability maximizing in the indifferent opponent condition while taking both agents' actual predictions during that trial block into account. Additional incentives to improve performance on the following block were provided by informing participants that reaching optimal performance (± three cents) would

[1] During the first ten trials of the experiment, each opponent randomly adopted one of three possible initial strategies: random response, probability matching, or probability maximizing.

double their payoff, whereas suboptimal performance would result in halved earnings on the subsequent trial block.

Parameter Estimation and Model Evaluation

We estimated parameters for each individual separately based on the models' accuracy in predicting the observed choice sequence one step ahead for each trial. That is, all models generate trial-by-trial choice probabilities for both response alternatives on the basis of subjects' prior decisions, their associated payoffs and the respective model's parameter values. Employing maximum likelihood estimation we searched for the set of parameters that maximized the summed log-likelihood of the predicted choice probabilities across trials given each participant's observed responses with an iterative particle swarm optimization (Kennedy & Eberhart, 1995). For each individual, optimization proceeded iteratively with a total of 24 particles, 23 of which started at random positions while the final particle started at the best parameter combination from the previous iteration. Optimization terminated once the model fit did not improve further for at least five successive iterations. The following parameter bounds constrained the optimization process: $\alpha \in [0,1]$ for the learning parameter, $c \in [0,1]$ for the transformed sensitivity θ, and $\beta \in [0,1]$ for the additional outsmarting parameter.

The final fit of each learning model was compared to a baseline statistical model which assumes constant and statistically independent choice probabilities across trials (see e.g., Yechiam & Busemeyer, 2005), and hence, accounts for the data without presuming any learning. The stationary probability of choosing the more profitable option pooled across all trials ($p_1, p_2 = 1 - p_1$) is the only free parameter in this baseline model and, to account for divergent model complexities, both learning models are evaluated by comparing differences in Bayesian Information Criterion (BIC; Schwarz, 1978) statistics between learning and baseline model (see e.g., Yechiam & Busemeyer, 2005). If a learning model is superior to the statistical baseline model, i.e. accurately describes how subjects adapt their choice behavior over time, positive ΔBIC values result from this model evaluation.

Results

Behavioral Data

The mean proportion of choices to the more profitable choice option for each block of 100 trials averaged across participants in the two experimental conditions is displayed in Figure 1. For inferential statistics, we conducted Bayesian analyses in addition to conventional methods of hypothesis testing to quantify evidence in favor of the null and alternative hypotheses (Wagenmakers, 2007). We assume equal plausibility for the null and alternative hypotheses a priori and report the posterior probability for the null hypothesis, denoted as $p_{\text{H0}}^{\text{Bayes}}$, associated with each effect. A mixed model ANOVA revealed a significant main effect of trial block ($F(2.37,113.8) = 27.9$, $p < .001$, $\eta_p^2 = .367$, $p_{\text{H0}}^{\text{Bayes}} = .00$)[2], with predictions closer to the respective optimal response strategy in the last compared to the first block of 100 trials for both groups. In the mimicry condition, subjects' choice behavior accurately approached optimal probability matching towards the last trial block ($M = .76$), whereas an indifferent competitor elicited decisions more in line with a probability maximizing strategy ($M = .90$). This adaptive divergence of learning processes is emphasized by a significant main effect of competitor type across all trial blocks ($F(1, 48) = 11.7$, $p = .001$, $\eta_p^2 = .195$, $p_{\text{H0}}^{\text{Bayes}} = .03$). The competition type by trial block interaction did not reach statistical significance, although the Bayesian evidence was ambiguous ($F(2.37, 113.8) = 2.77, p = .058$, $\eta_p^2 = .055$, $p_{\text{H0}}^{\text{Bayes}} = .66$).

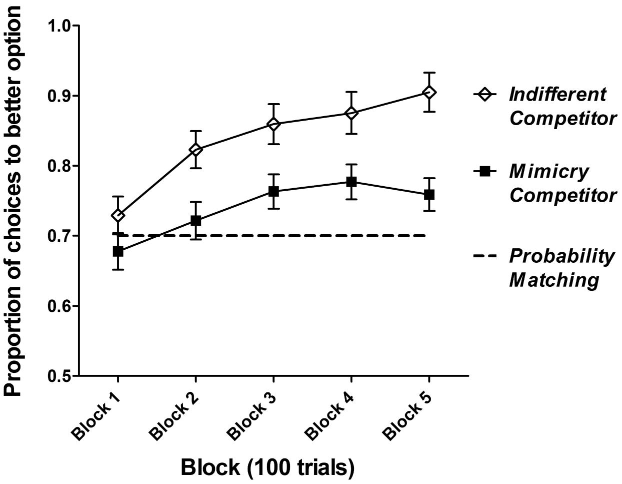

Figure 1: Mean ± standard error proportion of choices to the more profitable option averaged across trials and subjects.

Similar adaptive differences in choice behavior were also observed at an individual level, with high proportions of subjects in both conditions adopting the respective adequate rather than suboptimal strategy by the final trial block.

In sum, we have demonstrated that subjects are sensitive towards their competitors' decision strategies and modify their choices accordingly. The underlying psychological processes that lead to this adaptive divergence in strategy use, however, remain elusive from the behavioral data. Thus, we now turn to a computational modeling analysis to illuminate the determinants of emergent optimal choice behavior within competitive environments more holistically.

Modeling Data

The parameter estimates and ΔBIC values for the proposed learning models are compared in Table 1. The first learning model we examined defined decision utilities solely based on their associated monetary payoff and, judging by its

[2] Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(9) = 63.0, p < .001$), therefore degrees of freedom were corrected for both conventional and Bayesian analyses using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .593$).

Table 1: Mean and standard deviations (in parentheses) of parameter estimates and the difference in Bayesian Information Criterion (*ΔBIC*) between statistical baseline and specified model.

	Learning (α)		***Sensitivity (c)***		***Outsmarting (β)***		***ΔBIC***
	mimic	***indifferent***	***mimic***	***indifferent***	***mimic***	***indifferent***	
Monetary utility function	.07	.02	.16	.29	-	-	17.8 (31.3)
$u_t(i) = r_t$	(.21)	(.04)	(.20)	(.29)			
Competition utility function	.08	.07	.43	.28	***.85***	***.97****	18.7 (33.7)
$u_t(i) = [\beta \cdot g_t(i)] + [(1-\beta) \cdot s_t(i)]$	(.21)	(.14)	(.38)	(.22)	(.24)	(.05)	

*$p < .05$

clearly positive average Δ*BIC* score, accounts considerably better for the observed choice behavior than the stationary baseline model despite greater complexity. However, this basic utility model does not permit differentiation between the learning processes that lead to divergent choice behavior in the examined competitive environments, because estimated individual model parameters did not differ significantly between experimental groups, although the Bayesian evidence was ambiguous ($t(26.0) = 1.21$, $p = .238$, $p_{\mathrm{H0}}^{\mathrm{Bayes}} = .71$ for learning rates and $t(42.6) = -1.87$, $p = .068$, $p_{\mathrm{H0}}^{\mathrm{Bayes}} = .51$ for sensitivity constants).

The second learning model proposed above disentangles the two learning goals of choosing accurately, yet outsmarting the competitor by introducing an additional free parameter, *β*. The differential requirements of the two competitive environments are well represented by this additional outsmarting parameter, which was significantly smaller in the mimicry condition, indicating a tradeoff between betting on the more likely option and deviating from the opposing choice behavior, compared to the indifferent group, where opponent choices were to be disregarded ($t(26.4) = -2.44$, $p = .022$, $p_{\mathrm{H0}}^{\mathrm{Bayes}} = .27$). Parameter estimates for learning rate and sensitivity constant, again, did not differ between conditions, although the Bayesian evidence was ambiguous ($t(48) = .220$, $p = .827$, $p_{\mathrm{H0}}^{\mathrm{Bayes}} = .82$ and $t(38.6) = 1.65$, $p = .107$, $p_{\mathrm{H0}}^{\mathrm{Bayes}} = .59$, respectively). Although the more elaborate utility evaluation model sheds light on the processes underlying the observed divergence in choice behavior, the added complexity results in Δ*BIC* statistics not significantly better than those of the simpler utility model ($t(49) = -.613$, $p = .543$, $p_{\mathrm{H0}}^{\mathrm{Bayes}} = .88$). Thus, despite the conceptual promise and excellent parameter fit of the more complex model, overall, the simple monetary utility model is to be preferred for its parsimony.

Discussion

Qualitatively different competitive pressures in a binary prediction task result in adaptively divergent choice behavior on aggregate and individual learning levels. Under the influence of an indifferent opponent, resources should and were found to be exploited without consideration for the other agent's preferences, i.e. much like in classic individual binary prediction tasks, probability matching needed to be dismissed as an inferior strategy. By contrast, the presence of an imitative opponent necessitates response allocations proportional to outcome probabilities in order to maximize payoffs. In this context, we observed an adaptive tendency towards probability matching – i.e. probability maximizing was correctly rejected as an inferior strategy.

What drives this adaptive divergence of strategy adoption in these two competitive contexts? Our evaluation of learning models suggests that the observed adaptiveness of choice behavior largely results from differing learning goals with respect to opponent behavior: imitative competitors require consideration for strategies that influence and outsmart these agents, whereas indifferent opponents necessitate disregard for their choices when deciding on one's own course of action. Thus, conceptualizing opponent behavior as a key factor in the evaluation of choice utilities that is traded off against the desire to choose accurately disentangles these divergent requirements while providing a good approximation for observed behavior. Yet, when modeling individual data, the additional outsmarting parameter for each decision maker increased the complexity of the model beyond its explanatory potential as indicated by the Δ*BIC* score comparisons. Omitting the computational representation of opponent behavior from the model, however, results in parameter estimates that give little indication of the underlying learning processes prompting decision makers to respond adaptively to qualitatively different competitive pressures. At best, within this simpler model, divergent environmental requirements are somewhat reflected in marginally decreased sensitivities for evaluated choice propensities in the mimicry competitor condition, i.e. adoption of optimal probability matching is explained in terms of greater randomness in subject's choice behavior. Attributing the observed adaptiveness of strategy use in both contexts to differences in choice rule precision appears, however, conceptually implausible, because under the influence of an imitative competitor, participants are not less sensitive towards monetary rewards per se. On the contrary, we suggest that it is the added requisite to outmaneuver the opposing agent that fuels optimal matching in this context.

Consequently, to account for core learning processes that drive adaptive choice behavior within these competitive environments, an additional representation of opponent behavior is conceptually essential. To remedy potential disadvantages of added model complexity an interesting avenue for future research is to explore the suitability of hierarchical parameter estimation techniques, which may highlight the benefits of including an outsmarting parameter without introducing the downsides of overly complex models. The take-home message from this study is that learning to choose under uncertainty can indeed be steered by competition and thus proceed adaptively in situations where probability maximizing *or* matching is optimal.

Acknowledgments

The authors would like to thank Jörg Rieskamp and Lael Schooler for insightful suggestions on developing a model of competitive reinforcement learning. This research was supported by Australian Research Council grants to BRN (DP110100797; FT110100151), and a University of New South Wales International Postgraduate award (to CS).

References

Allee, W. C. (1978). *Animal aggregations: A study in general sociology*. New York: AMS Press.

Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, *88*(4), 848–881.

Gallistel, C. R. (1990). *The organization of learning* (1st MIT Press paperback). Cambridge, Mass: MIT Press.

Gigerenzer, G. (2000). *Adaptive thinking*. Oxford: Oxford University Press.

Godin, J.-G. J., & Keenleyside, M. H. A. (1984). Foraging on patchily distributed prey by a cichlid fish (Teleostei, Cichlidae): A test of the ideal free distribution theory. *Animal Behaviour*, *32*(1), 120–131.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461–482.

Harper, D. G. C. (1982). Competitive foraging in mallards: "Ideal free' ducks. *Animal Behaviour*, *30*(2), 575–584.

Janssen, C. P., & Gray, W. D. (2012). When, What, and How Much to Reward in Reinforcement Learning-Based Models of Cognition. *Cognitive Science*, *36*(2), 333–358.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *IEEE International Conference on Neural Networks Proceedings* (Vol. 4, pp. 1942–1948).

Keren, G. B., & Wagenaar, W. A. (1985). On the psychology of playing blackjack: Normative and descriptive considerations with implications for decision theory. *Journal of Experimental Psychology*, *114*(2), 133–158.

Rieskamp, J., & Otto, P. E. (2006). SSL: A Theory of How People Learn to Select Strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, *6*(2), 461–464.

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.

Singh, S., Lewis, R., & Barto, A. G. (2009). Where Do Rewards Come From? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2601–2606). Austin, TX: Cognitive Science Society.

Sutton, R. S. (1998). *Introduction to reinforcement learning*. Cambridge, Mass: MIT Press.

Todd, P. M., & Gigerenzer, G. (2007). Environments That Make Us Smart. *Current Directions in Psychological Science*, *16*(3), 167–171.

Vulkan, N. (2000). An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, *14*(1), 101–118.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.

Yechiam, E., & Busemeyer, J. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*, 387–402.

Yechiam, E., & Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. *Journal of Mathematical Psychology*, *51*(2), 75–84.

Appendix

Expected reward proportions are defined as the weighted average of all possible outcomes resulting from nature's move and both agents' choices. When two decision makers follow the same course of action, i.e. one imitates the other, the choice probabilities of both agents are identical. Given identical choice probabilities, the agents can either converge or diverge on a choice option, and thus expected rewards can be broken down into split and full payoffs while their sum amounts to the total expected payoff proportion:

$$r_{Split} = (p_c(H)^2 \cdot .7 + p_c(L)^2 \cdot .3)/2 \qquad (5)$$

$$r_{Full} = (p_{c1}(H) \cdot p_{c2}(L) \cdot .7) + (p_{c1}(L) \cdot p_{c2}(H) \cdot .3) \qquad (6)$$

For each decision maker, split reward proportions are computed as the joint probability of both agents choosing the same option ($p_c(i)^2$) weighted by the outcome contingencies (here, .7 and .3) and split by two; whereas full reward proportions can be expressed as the joint probability of both players choosing different options ($p_{c1}(i) \cdot p_{c2}(j)$) weighted by the outcome probabilities. Thus, total expected payoffs are maximized when both players probability match. For outcome probabilities of .7 and .3, for example, each player's maximal total expected reward proportion equals .395 (compared with .35 for probability maximizing).