A Reinforcement Learning Model of Bounded Optimal Strategy Learning

Xiuli Chen (xxc116@cs.bham.ac.uk)
School of Computer Science, The University of Birmingham
Edgbaston, Birmingham, B15 2TT, UK

Andrew Howes (A.Howes@cs.bham.ac.uk)
School of Computer Science, The University of Birmingham
Edgbaston, Birmingham, B15 2TT, UK

Abstract

In this paper we report a reinforcement learning model of how individuals learn the value of strategies for remembering. The model learns from experience about the changing speed and accuracy of memory strategies. The reward function was sensitive to the internal information processing constraints (limited working memory capacity) of the participants. In addition, because the value of strategies for remembering changed with practice, experience was discounted according to a recency-weighted function. The model was used to generate predictions of the behavioural data of 40 participants who were asked to copy appointment information from an email message to a calendar. The experience discounting parameter for a model of each individual participant was set so as to maximize the expected rewards for that participant. The predictions of this bounded optimal control model were compared with the observed data. The result suggests that people may be able to choose remembering strategies on the basis of optimally discounted past experience.

Keywords: bounded optimal; reinforcement learning; information processing bounds; memory constraints.

Introduction

Human beings are *bounded optimal* if they are able to maximize utility subject to the bounds imposed by their information processing capacities and their experience (Howes, Vera, Lewis and McCurdy 2004; Lewis, Vera and Howes, 2004; Howes, Lewis and Vera 2009). This paper reports progress towards a bounded optimal control theory of how people perform simple tasks that make use of memory. The model uses reinforcement learning to acquire optimal strategies given bounds imposed by short-term memory and experience. It therefore represents an example of a class of models that harness both the rigour of optimisation and theories of the bounds on human information processing (Anderson et al. 2004). The model also represents a departure from theories of unbounded optimisation (Griffiths & Tenenbaum 2006; Griffiths, Kemp and Tenenbaum 2008) and descriptive theories of bounds.

The model reported in the current paper captures what people choose to do given experience of the behavioural consequences of tasks that required memory. For example, when reading and writing a telephone number a person may choose to read the whole number, store it in memory, and then write it out. Alternatively he/she may choose to read the number 3 digits at a time and write out each 3-digit block before reading the next. There are many strategies but each has potentially different performance characteristics: Some might be fast but generate many errors, others relatively slow but reliable. Tasks such as these have been investigated by Gray, Simms, Fu and Schoelles (2006). Gray et al. used the Blocks World task to study the choices that people make about what to remember. The participants were required to reproduce patterns of coloured blocks from a Target window to a Workspace window. For example, there might be 8 different coloured blocks which were positioned randomly in a 4x4 grid. The number of blocks encoded by a participant on each visit was regarded as corresponding to a strategy. Gray et al. demonstrated that participants were able to adapt their choices of strategy to the cost/benefit structure of the environment given experience.

More recently, Howes et al. (submitted) employed a similar task, called the Email-Calendar Copy task, in which the participants were required to copy the appointment information from an email interface to a calendar. The results suggested not only that participants were able to adapt their choice of strategy, as demonstrated by Gray et al. (2006) but also that many would end up preferring the optimal strategy given their learned knowledge. The reinforcement learning model reported in the current paper is a model of the results of Howes et al. (submitted). Unlike with many previous reinforcement learning models, including those of Gray et al. (2006), the current model parameters were chosen so as to maximize utility, not so as to maximize fit. The predictions of the model were then compared to the participants' behaviours. The results suggest that when people learn which strategy to use through reinforcement learning, they may do so by using optimal discounting of past experience.

The remaining paper is organized as follows. The task is introduced in the next section and is followed by a description of the model, called the bounded **Optimal Discounting (OD) model**. Subsequently, the model results are presented, followed by a comparison between the current model predictions and those predicted by an alternative model in which the individual models use the same discounting parameter, which is called **Non-optimal Discounting (ND) model** .

The Task

The modeled data was acquired from the experiment reported by Howes et al. (submitted). The participants were required to copy appointment information from an email interface to a calendar. Appointments were presented in trials. On each trial, participants were asked to view various numbers of appointments on the email window one by one, ranging from 3 to 9. Since the first appointment was always at 09:00 AM and these appointments were always one hour apart and in sequence, only the names and the order they were presented need to be remembered. Once the last appointment was shown, the 'OK' button on the email window enabled the participants to go to the calendar window, with the email window disappearing, and copy these appointments across by typing these names in the time slots. Once they were satisfied with their copy and clicked the 'Finish' button, they would receive feedback about the number of appointments correctly copied and highlighted in red any slots incorrectly completed.

An important difference between the studies of Howes et al. (submitted) and of Gray et al. (2006) is that the Howes et al. study was designed with two-phases, a no-choice phase followed by a choice phase. In the no-choice phase, the *strategy* that participants adopted on each trial was assigned by the system (the number of appointments that participants were required to view before copying across was regarded as a *strategy*, ranging from 3 to 9). During this phase, they were asked to copy 100 correct appointments (only correctly copied items were counted in the target total items), and the strategies (3, 4, 5, 6, 7, 8 and 9) appeared almost evenly. The reason to do so was to force the participants to explore across the strategy space so that it allows us to empirically measure their performance over the strategies. After this phase, entering in Choice-phase, the participants were required to copy 200 appointments correctly by selecting their own preferred strategies on each trial. In addition, participants were asked to minimize the total time taken for the task and as they had to copy a target number of correct items, they were effectively asked to optimize the speed/accuracy trade-off. Therefore, the utility of each strategy was defined in term of *reward rate,* which was defined as the rate of successful copies.

The Bounded Optimal Discounting (OD) Model

The purpose of the model is to explain strategy choice on simple remembering tasks. As we have said, rather than maximizing the fit of the model to the data, a key feature of the model is that remembering strategies, and the experience discounting parameter, are chosen so as to maximize utility.

The remembering strategy space consists of strategies for remembering 1 to 9 items on each visit to the calendar. The choice of the discounting parameter, named *StepSize*, has consequences for the weight given to a reward when estimating the future utility of a remembering strategy. In our model, the discounted parameter that is used to update the trial-by-trial strategy value estimates is set so as to optimize the overall utility of the model for each individual.

Detailed Description of Optimal Discounting (OD) model

RL is concerned with learning to obtain rewards or avoid punishments by trial and error (Sutton & Barto 1998; Daw & Frank, 2009; Cohen, 2008). It has been used to understand how iterated rewards and punishments (experience) determine choice behavior in various situations. In particular, how the structure, amount, hierarchy etc. of the observed experience relate to the learning results has attracted increasing attention (Botvinick & Barto 2009). A reinforcement learning model with strategy-utility updating based on recency weighted experience is used in our analysis.

The model is defined by three parameters, [**S, R, E**], and strategy-value estimation updating rules. **S** is the strategy space, $\mathbf{S= \{S_1, S_2, \ldots Si, \ldots S_n\}}$. The strategy taken on trial **t** is denoted **S(t)**. Once the strategy has been selected, the environment would give reward from the reward set **R**, $\mathbf{R \in [0, 1]}$. The reward following the strategy S_i on trial **t** is denoted as $\mathbf{r_i(t)}$. In this learning problem, each strategy has an expected or mean reward given that that strategy is selected, called the true (actual) *value* of this strategy. To measure the utility of the strategies trial-by-trial, the model uses estimated values acquired through experience. Specifically, on each trial **t**, $\mathbf{t \in [1,2\ldots]}$, the model updates an estimate vector, $\mathbf{E(t)=\{E_1(t), E_2(t), \ldots, E_i(t), \ldots, E_n(t)\}}$, where $\mathbf{E_i(t)}$ is the estimate of strategy value of $\mathbf{S_i}$ on trial **t**. The initial estimated value of each strategy is 0, i.e. $\mathbf{E_i(0)=0}$, where $\mathbf{i \in [1,2\ldots , n]}$. In addition, because the values of remembering strategies are non-stationary, due to practice, a discounting technique is applied to the experience when estimating the strategy-value. Specifically, as people practice a strategy they improve. This process of improvement means that for any pair of strategies i and j, the relationship between their true values at trial t, e.g. $\mathbf{V_i(t)}$, > $\mathbf{V_j(t)}$, will not necessarily hold after an increment in the practice of i, the practice of j, or both. Therefore, in order to track this non-stationary learning environment, more recent experience might deserve to be weighted more heavily than temporally distant experience. Here we adopt one of the most popular ways to achieve experience discounting, called the *exponential, recency-weighted discounting.* Specifically, if a strategy has been chosen **k** times before, yielding rewards $\mathbf{r_1, r_2 \ldots r_k}$, then the value of this strategy is estimated to be

$$\mathbf{E_k = \sum_{i=1}^{k} \left(\frac{c(1-c)^{k-i}}{\sum_{i=1}^{k} c(1-c)^{k-i}} \right) \times r_i} \qquad (1)$$

Where **c** ($\mathbf{c \in (0,1]}$) is the discounting parameter, called *StepSize*, which determines the weighting of previously received rewards. The weight given to a reward r_i depends on how many rewards previously, k-i, it was observed. As (1-c) is always less than 1, the weight given to r_i decreases as the number of intervening rewards increases. In fact, the weight decays exponentially according to the exponent on 1-c. The higher the value of the *StepSize*, then the more recent rewards will contribute to the estimate relative to distant rewards. Figure 1 below gives the weight distributions of

$$\mathbf{f(c)} = \frac{\mathbf{c\,(1{-}c)^{k-i}}}{\sum_{i=1}^{k}\mathbf{c\,(1{-}c)^{k-i}}} \qquad (2)$$

with k=8, i=1, 2, … 8, and five sets of c, [0.1, 0.3, 0.5, 0.7, and 0.9]. As you can see, the line of c=0.1(the red one) is much more flat than the line of c=0.9 (the green one), which means that the relative distant rewards, like $r_{5\,,}$ $r_{6,}$ under the c=0.1 model contribute more to estimate the future utility than they do under the model with c=0.9 in which the estimation mostly relies on two latest rewards r_7, r_8.

In the **OD** model the value of the *StepSize* parameter was chosen so as to optimize utility for each individual. Specifically, given the means of the estimated strategy values above and a specific *StepSize*, then on each trial, the strategy with the highest estimate, i.e. the greedy action, is taken as the prediction of the participant's behaviour. For these predicted strategies, the model also gives predictions of their rewards. On each trial, the mean of the rewards received by the predicted strategy is regarded as the predicted reward of this strategy. Therefore, for each set of the *StepSize* we get a set of predicted strategies and rewards for each participant. We find a *StepSize* that generates maximal overall reward for each participant.

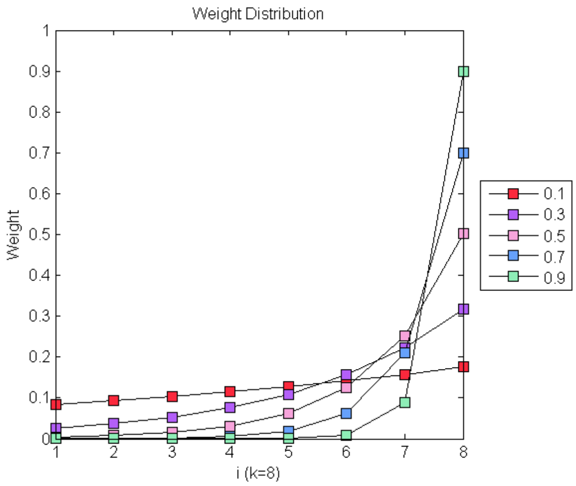

Figure 1: The weight distributions.

Alternative Model

In order to test the **OD** model we compared it with a model in which the *StepSize* is set to be 0.1 for all the participants. This value offers very little discounting (Figure 1). Specifically, on each trial, the values of the strategies which are selected for k times with rewards $\mathbf{r_{1},\ r_{2}\ldots r_{k}}$, are estimated according to the equation (1) with c=0.1. In other words, for this model, there are two key features. First, according to the weight distribution with c=0.1 in Figure 1 you can see that the weights put on the experienced rewards differ only a little, which means that previous rewards almost equally contribute to the future utility estimation. Second, the same parameter value is used for all the participants. As with the OD model, the greedy action on each trial of the choice phase is predicted to be the participants' behavior, and on each trial, the mean of the rewards received by the predicted action is regarded as the predicted reward of this action. We call this model the **Non-optimal Discounting (ND) model**, i.e. the model with a fixed low- discounting parameter for all participants.

Predictions of the models

Both models, **OD** and **ND**, predict trial-by-trial individual participant strategy selections on the basis of the strategy-value estimates. In addition, they predict the rewards following the predicted actions, so that we could find a predicted action set that maintains the maximal expected reward. As we have said, for **OD**, the weights given to the rewards are adjusted by setting the discounting parameter **c** to a value that optimizes expected reward for each participant. For **ND**, the parameter is set to be 0.1 for all the participants. Comparison between **OD** and **ND** model allows us to test the assumption that people adapt the discounting parameter so that it is optimal given the constraints imposed by practice. If **OD** makes significantly better predictions than **ND** then we have evidence that participants discounted their previous experience given the expected effects of practice on strategy value.

Consequentially, for both models we obtained the predicted actions and reward rates on each trial. Despite the fact that neither model is fitted to the data we expect **OD** to offer significantly better predictions than **ND**.

Results

Overall Performance over the Strategy Space

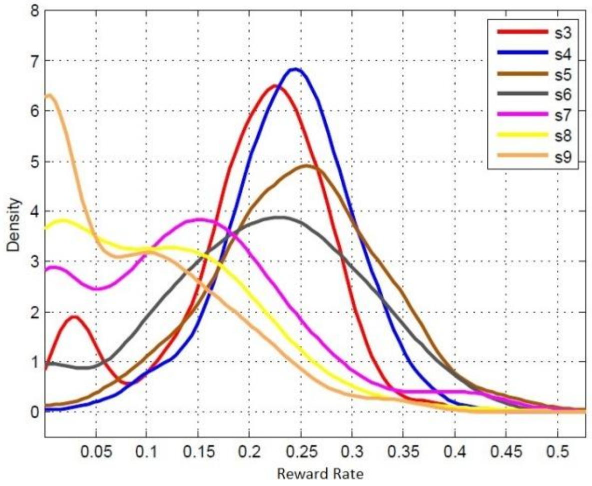

Figure 2: the probability densities of the reward rate for each strategy over all the participants.

For each participant and each trial, the following experimental data was recorded: selected strategy (one of 3, 4, 5, 6, 7, 8 or 9 items, including the strategies assigned by the system in the no-choice phase and the strategies chosen by the participants in the choice phase), the number of correctly copied items, and the trial duration. The *reward rate* of the selected strategy is computed as the number of items correctly copied at a trial over the trial duration. Figure 2 (above) gives the overall measurement of each strategy's performance over all the 40 participants during the experiment. As shown in the figure, strategies 3, 4, 5 are the three most effective strategies across participants (Mean=5.6773, SD=1.8505, Mode=5). It is also evident that some of the strategies have bimodal densities, reflecting the low reward rates associated with error trials.

Descriptive Results

First consider the predictions of the **OD** model. As mentioned above, an OD model with a discounting parameter *StepSize* that maximizes the sum of predicted rewards over the choice phase was found for each participant. In Figure 3 (below), each panel represents trial-by-trial value estimates for a participant. X-axis represents trials; Y-axis is the strategies' value estimations calculated by the **OD** model trial by trial. Each strategy is represented by a different colour, as shown in the legend on the right side of the figure. To the left side of the vertical black line is the no-choice phase; on the right side is the choice phase. The participant strategy on a trial is represented by a black circle (including the strategies assigned by the system in the No-Choice phase and the strategies chosen by the participants as their preferences in the Choice phase). The title of each panel includes information about the participant number and the *StepSize* found for the participant. Participants 15, 19, 7, 8 were selected to demonstrate the diversity of the individual performance. For comparison we divide the participants into three groups.

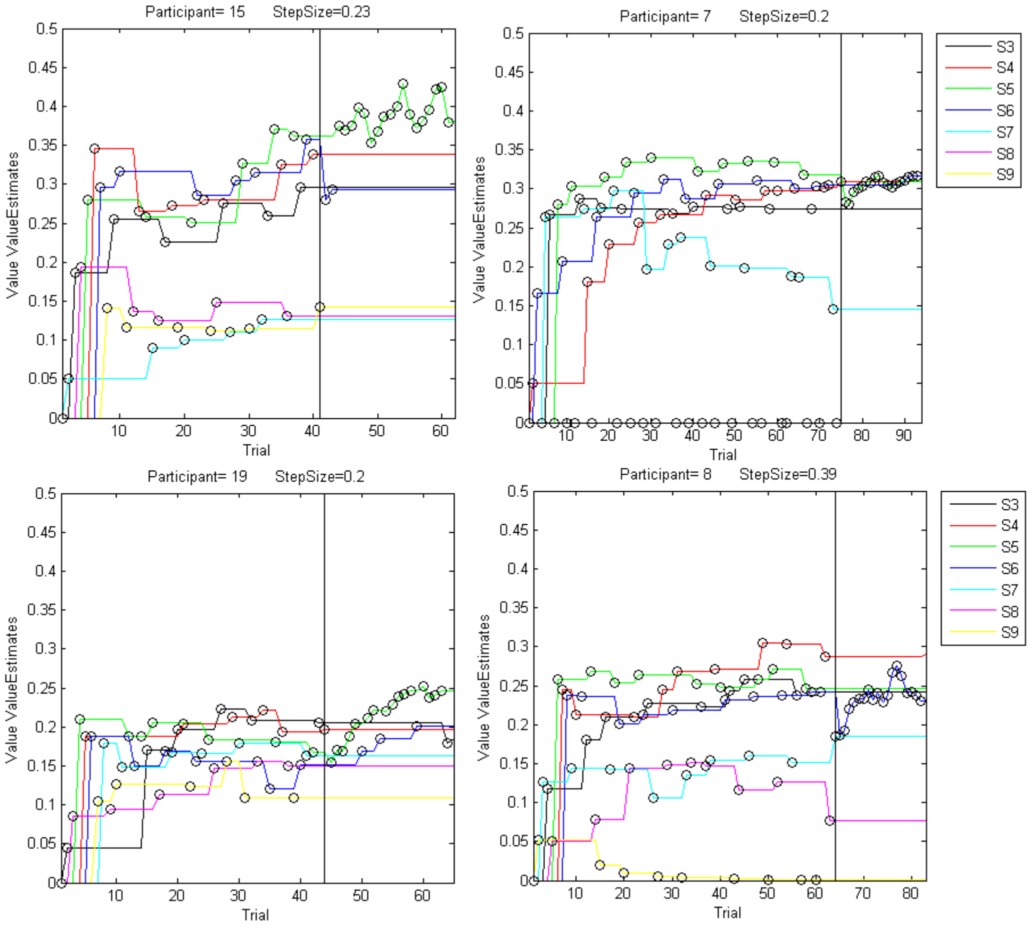

Figure 3: **OD** model predictions. X-axis represents trials; y-axis is the value estimates for the strategies calculated with the **OD** model. Each strategy is represented by a different colour, as shown in the legend on the right. To the left side of the vertical black line is the no-choice phase, on the right side is the choice phase. The selected strategy on a trial is represented by a black circle (including the strategies assigned by the system in the No-choice phase and the strategies chosen by the participants in the Choice phase). The title of each panel includes the information of the participant number and the *StepSize* found for the participant.

Group 1: *The best strategy was selected on the majority of trials in the choice phase, such as participants 21 and 14.*

Specifically, for participant 15 (top left panel), the strategy S5 became the best one (with the highest value estimate) by the end of the no-choice phase, and the participant used it on most trials in the choice phase. While for participant 19 (bottom left panel), the strategy S5 is not the best at the beginning of the choice phase, but its performance improved with practice, became best, and was chosen by the participant at the later stage of the choice phase. For the **OD** model, 27 of the 40 participants exhibited a pattern that was either consistent with participant 40 or 20. (*StepSize* was found between 0.03 and 0.82). For the **ND** model, 22 out of 40 participants behave in this way.

Group 2: *There is no clear bounded optimal strategy in most trials of the choice phase, e.g. participant 7.*

For some participants such as participant 7 (top right panel), there are several best strategies (in this case, S4, S5 and S6) with, informally, close value estimates, or it is the case that the best strategy frequently changes during the choice phase. Therefore, many strategies appear to have the highest reward and it is rational to keep exploring through the choice phase. Overall, for the **OD,** 8 out of 40 participants were predicted to be in this group, while 9 out of 40 for the **ND** model.

Group 3: *There was a clear best strategy predicted, but the participant did not end up choosing it, e.g. participant 8.*

From the beginning of the choice phase, S4 was a clear best strategy for participant 8 (bottom right), but the participant chose the strategy S6, which was unlikely to be the highest reward strategy. Overall 5 of the 40 participants behave in this pattern according to the **OD** model. For the **ND** model, 9 out of 40 participants are in this group.

Model Comparison

We computed the Root Mean Square Error (RMSE) between the strategies predicted by the model and the observed participant behaviours. The Lower RMSE, the better the model prediction. For the **ND** model, RMSE between predicted and observed actions in the choice phase is 1.2845, while it is 1.1539 for the **OD** model. In addition, we calculated RMSE between the received rewards and the predicted rewards for these two models, 0.0782 and 0.0703 for **ND** model and **OD** respectively.

We computed t-tests on the Mean Squared Errors (MSE) to determine which of these two models offered better predictions on the strategy during the choice phase. A paired right-tailed t-test between **ND** model and **OD** model indicated that the **OD** model, with the discounting parameter that maximises the expected reward rate is able to offer significantly better predictions of strategy choice (t(39)=1.80, p=0.0396).

Discussion

The results support the hypothesis that a model that makes bounded optimal use of internal resource (memory and experience of reward) so as to select strategies for remembering is able to predict the majority of participant choices. In particular,

(1) For the **OD** model, a discounting parameter, *StepSize*, was used to control the weights put on the rewards received by the strategies when estimating the values of the strategies for predictions of subsequent behaviour. The **OD** model with the *StepSize* that maximized the expected rewards for each individual participant offered a significantly better prediction of the observed data than the **ND** model, which weighted the received rewards with a fixed, minimal, parameter value of 0.1 to estimate the value of the strategies for all participants.

(2) The *StepSize* that maximized the expected reward for the participants had a large range, ranging from 0.03 to 0.82. This may reflect the ability of participants to optimally adjust learning parameters to reflect meta-knowledge about the effects of their own practice on skill.

General Discussion

According to a number of studies and models, memory bounds human performance in many complex tasks, e.g. reasoning, comprehension, and learning (Cowan, 2005; Vaughan & Herrnstein, 1987). The reported study suggests that people can make bounded optimal use of memory in an everyday interactive task (copying information from email messages). In addition, people are able to strategically adjust learning parameters in response to estimates of expected reward that are non-stationary because of practice of a cognitive skill. It appears that people are able to do as well as they do on remembering tasks by selecting optimal strategies according to the cost/benefit structure of their own discounted experience of practice.

Our finding in favour of optimal strategies was not supported by the data from every individual participant. For example, participant 11 was still highly exploratory in choice phase and was not predicted by the model. However, the findings do suggest that a model that uses an optimal discounting parameter *StepSize* (**OD** model) does make better predictions than a model in which a fixed discounting parameter is used to predict all participants and most rewards information from the experience are used almost equally (**ND** model with c=0.1).

Further tests of the model are required to determine, for example, how well the OD model does relative to the best-fitting model, where the best fitting model adjusts *StepSize* so as to fit the data. It is inevitable that the best-fitting model will be at least as good as OD but any gap between how well the two models correspond to the data will tell us something about how much variance is unexplained by OD.

There was also evidence that some participants selected strategies that were not optimal in the early parts of the choice phase, but that with practice were improved and by the end they were generating the highest rewards. This fact is consistent with the observation that the learning environment was non-stationary because of the acquisition of knowledge through practice. There are many studies that focus on the improvement of strategies with practice but in this paper our focus has instead been on how choices are made between strategies given that, through the effects of practice, strategies have non-stationary utility. Our starting point is the assumption that an estimate of the future utility of a strategy can be based on previous experience but that in the non-stationary environment construed by practice, it is valuable to discount the past so that more recent experience is weighted more heavily than temporally distant experience.

Conclusion

The paper provides quantitative evidence for the hypothesis that people are bounded optimal when learning to choose strategies that improve with practice. They appear to be able to manage their internal resource and learning strategies so as to maximize performance against an externally imposed payoff function.

Acknowledgments

The work was funded by a PhD scholarship awarded by the Human-Computer Interaction Centre, School of Computer Science, University of Birmingham.

References

Anderson, M. C., Ochsner, K.N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S.W., Glover, G.H., Gabrieli, J.D.E. (2004). Neural systems underlying the suppression of unwanted memories. *Science*, *303*, 232-235.

Botvinick, M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262-280.

Cohen, M. X. (2008). Neurocomputational mechanisms of reinforcement-guided learning in humans: A review. *Cognitive, Affective & Behavioral Neuroscience, 8*, 113-125.

Cowan, N. (2005). *Working memory capacity*. Hove, East Sussex, UK: Psychology Press.

Daw, N. D., & Frank, M. J. (2009). Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition*, 113, 259-261.

Gray, W. D., Simms, 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, 461-482.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition, *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.

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

Howes, A., Lewis, R.L. & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review, 116*, 4, 717-751.

Howes, A., Vera, A., Lewis, R.L. McCurdy, M. (2004). Cognitive Constraint Modeling: A formal approach to supporting reasoning about behavior. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, Chicago.

Howes, A., Duggan, G., Tseng, Y-C., Kalidindi, K. & Lewis, R.L. (submitted). A bounded optimal theory of short-term remembering. Paper submitted for journal publication.

Lewis, R.L., Howes, A., Vera, A. (2004). A constraint-based approach to understanding the composition of skill. *International Conference on Cognitive Modeling*, Pittsburgh, 2004.

Vaughan, W., Jr. & Herrnstein, R.J. (1987). Stability, melioration, and natural selection. In L. Green & J.H. Kagel (Eds.), *Advances in Behavioral Economics*, Vol. 1 (pp. 185–215). Norwood, NJ: Ablex.