Learning near-optimal search in a minimal explore/exploit task

Ke Sang (kesang@indiana.edu), Peter M. Todd (pmtodd), Robert L. Goldstone (rgoldsto)
Cognitive Science Program and Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

How well do people search an environment for non-depleting resources of different quality, where it is necessary to switch between exploring for new resources and exploiting those already found? Employing a simple card selection task to study exploitation and exploration, we find that the total resources accrued, the number of switches between exploring and exploiting, and the number of trials until stable exploitation becomes more similar to those of the optimal strategy as experience increases across searches. Subjects learned to adjust their effective (implicit) thresholds for exploitation toward the optimal threshold over 30 searches. Those implicit thresholds decrease over turns within each search, just as the optimal threshold does, but subjects' explicitly stated exploitation threshold *increases* over turns. Nonetheless, both the explicit and learned implicit thresholds produced performance close to optimal.

Keywords: exploration; exploitation; explore/exploit tradeoff; optimal search; threshold strategy.

Introduction

Search is a ubiquitous requirement of everyday life. Scientists need to search for information to help their research; web users use search engines like Google to get whatever they are interested in from the internet; companies search for the best candidates for their job openings; consumers searching in supermarkets with hundreds of brands of candies have to decide if they have found one that is good enough or if they should explore to find something even tastier.

In many real life situations, to search (or explore) or to stop searching (and exploit the fruits of the search) is a key issue for making better decisions. Organisms have to make tradeoffs between exploration and exploitation so as to improve their success in the environment. Consider a honeybee searching for nectar in flowers. Suppose the honeybee has visited a particular plant and found most of the nectar in its flowers. The bee must decide whether it is worth spending more time to find still more nectar on this plant, exploiting it further, or whether it would be better off leaving this plant and exploring to look for another. Staying too long on the flowers of this plant is wasteful, and the bee should move to another plant with higher initial rate of nectar supply; however, leaving that initial flower plant too early is also suboptimal because travelling between resource patches will cost time and energy, and there is uncertainty about the resource levels of flowers that have not yet been visited. To maximize intake of nectar, the bee needs a decision rule that balances exploration of new resource sites with exploitation of known resource sites (Charnov, 1976).

The same tradeoff between exploiting what you already have and exploring further to find something preferable applies to humans. For instance, should you take the parking space you have just found or keep driving closer to your destination hoping to find a better one? Should you stick with your current job, or partner, or brand of coffee, or explore further to see if there are better options to be found?

Many researchers have focused on aspects of exploration versus exploitation. Optimal decision mechanisms and heuristic rules of thumb have been proposed to model when animals leave patches to find new ones (Charnov, 1976; Bell, 1991; Livoreil & Giraldeau, 1997; Wajnberg, Fauvegue, & Pons, 2000). Mathematicians have studied optimal stopping problems where the task is to decide when to stop the exploration phase of search and exploit a particular chosen option; Ferguson (1989) reviews work on one well-known form of this task, the so-called Secretary Problem. Todd and Miller (1999) applied this kind of framework to the problem of searching for a mate, studying the simple heuristics that could work well to stop exploratory search once an appropriate partner was encountered, and Beckage, Todd, Penke, and Asendorpf (2009) found evidence of use of such rules by people searching for mates at speed-dating events. Lee (2006) developed Hierarchical Bayesian models to account for human decision making on an optimal stopping problem.

Different resource types and environmental structures call for different search strategies. Thus, how well humans perform in experiments involving the exploration/ exploitation tradeoff depends on the task details, which influence not only optimal search strategies, but also the actual strategies employed by subjects. In this paper we focus on search behavior in a resource-accumulation setting, in which individuals make a series of decisions as to whether to explore to find a new resource or exploit a previously-encountered one, accumulating value from both newly-found and previously-discovered, currently-exploited resources as they search.

Search Task

In the experiment, individuals had to accrue as many points from cards as possible over a 20-turn game. At each turn, a subject could either explore by flipping over a card with unknown points from a card deck, or exploit a card already uncovered by selecting it from a computer screen. With this accumulation of resources (e.g. points) during both exploration and exploitation and the ability to return to previously-found items, this search task resembles a non-competitive foraging task with non-depleting resources.

Note that these task settings differ from the classic Secretary Problem and the widely studied patch-foraging problem. Compared to the Secretary Problem, individuals in our experiment have knowledge of the outcome distribution (card values are uniformly distributed from 1 to 99). They can switch from exploitation back to exploration (even though this is never done by the optimal strategy), whereas the Secretary Problem involves searching (exploring) until a single option is chosen (exploited). Individuals are also able to go back and exploit previous items, and they receive points in both exploration and exploitation phases, whereas the Secretary Problem payoff is determined solely by the final choice made. In a typical patch-foraging problem, foragers usually do not know the distribution of resources in patches, exploring between patches has costs, and exploiting a patch makes its value go down over time (depleting resources), so that foragers usually do go back and exploit previously-found patches even though they could.

Many possible rules could describe subjects' behavior in our search task. These include intertia-based rules (subjects have a tendency to repeat the previous action, be it exploration or exploitation), impatience-based rules (after some number of turns doing one action, individuals lose patience and switch to the alternative action), and threshold rules (switch from explore to exploit when a value above a particular threshold is found). We focus here on threshold rules, in part because that is the form of the optimal strategy.

Optimal Strategy

To judge how well subjects perform, it is useful to understand the optimal strategy for the given task settings in our experiment. The optimal strategy is to use a decreasing threshold, switching from exploration to exploitation whenever the best card seen so far exceeds the current threshold level. According to the optimal strategy, the decreasing threshold curve only depends on the range of card values (highest and lowest) and the total number of turns in one search game.

Let H denote the highest possible value for a card, L denote the lowest possible value, N denote the total number of turns in one game, n denote the current turn within the game, and d_n denote the optimal threshold value for the n^{th} turn.

Also let:

$A = (N-n)\cdot(H^2+H)$,

$B = (H+L)\cdot(H-L+1)$,

$C = (N-n)\cdot(2H+1) + 2(H-L+1)$,

$$\text{then } d_n = \begin{cases} \dfrac{C - \sqrt{C^2 - 4(N-n)(A+B)}}{2(N-n)}; & \text{when } n < N. \\ \dfrac{H+L}{2}; & \text{when } n = N. \end{cases}$$

What would the threshold curve look like? In our experiment, H=99, L=1, and N=20. The threshold curve for these values is plotted in Figure 1.

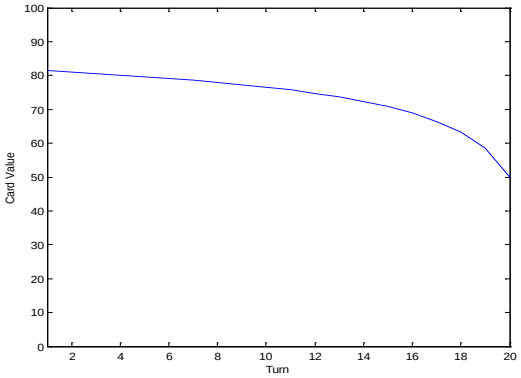

Figure 1: Optimal threshold curve. Over turns, the optimal threshold for exploiting the highest drawn card falls from about 80 on turn 1 to 50 on turn 20.

Experiment

191 subjects were recruited for the experiment from the Indiana University psychology student subject pool. They were told that their goal was to accumulate as many points as possible in each search game, by flipping over cards whenever they wanted or taking the points from cards already exposed on the screen. Subjects were also informed that the point values for cards ranged from 1 to 99, with equal probability (i.e., card values were uniformly distributed between 1 and 99), selected with replacement.

In the experiment, a turn refers to one time of either exploration or exploitation, and every trial contains 20 turns. The interface for the experiment is shown in Figure 2. Every card had its value displayed on it. In the first of the 20 turns, the subject must explore, flipping over the top card on the deck. After seeing its value and having that added to their accumulating points, subjects could do either of two actions on the second turn (and all subsequent turns): select a new card from the deck (exploring), or select one of the cards that he/she had already turned over (exploiting). The screen displayed the number of turns taken, the total points obtained thus far for this trial, and the highest card value seen so far in the trial (by showing that card's point value in red on the card, while all other cards were shown in green).

For example, in Figure 2, four cards have been taken from the card deck, with the first three values in green while the highest card value, 91, is in a larger red font. The screen shows that the number of turns taken thus far is 15, there are 5 turns left, and the total points so far for this trial is 1245. The number of points received by the subject on each turn in this trial is also listed beside the deck. On this 16th turn, the subject should decide whether he/she wants to exploit the highest value 91 again, as they have done for the previous 12 turns, or explore the deck hoping for a higher card value.

After each of the 30 independent trials, subjects were told the points they received and the points that the optimal strategy would have earned. After finishing all 30 trials, subjects reported their explicit threshold—the minimum value of the maximum card seen so far that would lead them

to exploit that value rather than explore by flipping over a new card from the deck, for turns 2, 5, 9, 13, 17, and 20.

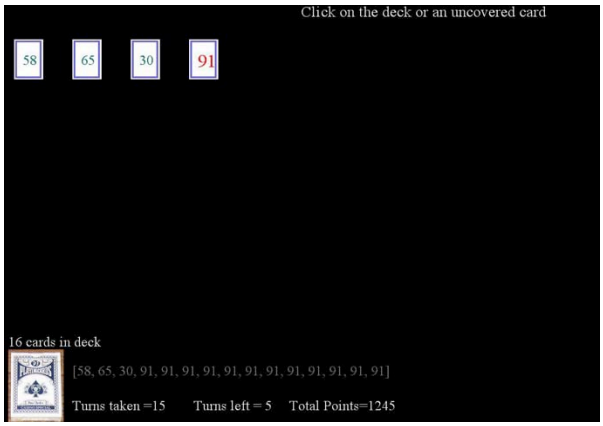

Figure 2: The interface of the experiment. The face-down card in the lower-left corner represents the deck of unknown cards, while the four cards in the upper portion of the screen represent turned-over cards.

Results

Across all of the turns taken by all subjects (191·30·20 = 114600 turns), there was 73.3% exploitation and 26.7% exploration. For the optimal strategy, there is more exploitation: 81%. Subjects' mean total points per 20-turn trial was 1528 (*SD* 266); for optimal, it was 1601.

Switch and Exploitation

The optimal strategy dictates that there would be at most one switch from exploration to exploitation per 20 turn trial—whenever the highest card seen so far exceeds the current threshold level. Subjects, by contrast, might switch back from exploitation to exploration for many reasons, including intrinsic randomness, boredom, or changing strategies over time. And then as the end of the trial approaches, they may well switch to exploitation again to take advantage of previously found high values. The data indicates that subjects switch between exploration and exploitation a mean of 1.83 times per trial.

In general, after some point subjects switch to exploitation and only exploit for the rest of the turns until the end of the trial. The turn where this continuing exploitation begins depends on the search strategy used. For example, a strategy with a constant threshold of 90 would lead to a later mean switch point than the optimal strategy does, because cards exceeding this high threshold are less common than cards exceeding the decreasing optimal threshold. The mean of the starting turn for this continuing exploitation is 7.35 across all subjects. We also simulated data for the same number (191*30 = 5730) of trials following the optimal strategy, and found the mean starting turn for continuing exploitation to be 5.14. Accordingly, people continue exploring for longer than optimal, but only by about two turns. Figure 3 shows the frequency distributions of these starting-final-exploitation turns for both the actual data and the optimal strategy. Compared to the optimal strategy, the distribution of the actual data has a long fat tail, which means that some subjects explored even until the very end and did not exploit a high value when they found it (which is likely—if someone explores for 10 turns, the probability that he/she will see a value larger than 80 is about 90%).

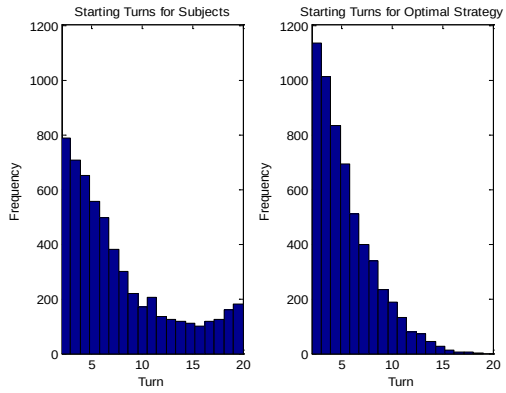

Figure 3: Frequency distributions of starting turns for final exploitation phase, for subjects (left) and optimal strategy (right).

Explicit and Implicit Thresholds

At the end of the experiment, we asked subjects to declare the minimum card value that they would have been satisfied with, and hence stop exploring and instead exploit this card for the rest of the turns. We asked them to disclose this value for turns 2, 5, 9, 13, 17, and 20. These values can be treated as indicating subjects' explicit thresholds; they are plotted in Figure 4, linearly interpolated. Generally speaking, this is an increasing curve, which moves in the opposite direction of the optimal threshold over turns.

At the individual level, we categorized subjects into four different types according to the trends of their explicit thresholds. If the reported thresholds at those 6 turns remained the same, subjects were classified as 'Constant'; if the values increased at least once and never decreased, subjects were classified as 'Increasing'; if the values decreased at least once and never increased, subjects were classified as 'Decreasing'; otherwise, they were labeled as 'Mixed'. Among 188 subjects (3 were excluded due to incomplete questionnaires), 71 subjects were Increasing, 47 were Decreasing, 19 were Constant, and 51 were Mixed. Not only is the general trend of the mean explicit threshold increasing over turns, but there are also far more subjects classified as individually 'Increasing' than 'Decreasing'.

As mentioned above, we focus on the threshold rule that subjects may use. In addition to subjects' explicit thresholds, we also analyzed the implicit thresholds that underlie their actual actions in the experiment. To estimate subjects' implicit thresholds, one way is to treat the implicit thresholds at different turns as parameters of cognitive

models and use the maximum likelihood estimation (MLE) method to estimate them. Here we propose two models, both of which have a stepwise threshold. The reason for using stepwise rather than continuous threshold models is that we want the estimates of the implicit thresholds to match up with the 6 separate explicitly reported thresholds.

Model A has 6 parameters, each representing a part of a stepwise threshold. Let T_i ($1 \le i \le 6$) be the 6 parameters; then T_1-T_6 respectively represent the thresholds that apply across turns 1-2, 3-5, 6-9, 10-13, 14-17 and 18-20. For each of these ranges of turns and corresponding T_i the model is:

$$Pr(explore) = \frac{1}{1 + e^{-0.1(T_i - Max)}}$$

Pr(explore) is the probability of exploration on the current turn, *Max* is the highest card value seen (on the table) before the current turn, and T_i has a range from 1 to 99.

Model B is nested with Model A, but has another free parameter, the sensitivity parameter *s*. It is a positive value that reflects how strongly the subject follows this threshold rule—if *s* is large, then subjects usually make a choice that is consistent with the threshold T_i, and if *s* is small, there can be a lot of randomness in the subject's choices. The model at each step for T_i is:

$$Pr(explore) = \frac{1}{1 + e^{-s(T_i - Max)}}$$

We used MLE to estimate parameter values of the two models for each individual. To select the model describing the data better, the Bayesian Information Criterion (BIC) was used to compare models. Because most of the parameter estimations are negatively skewed, here we chose the median, rather than mean, of the index BIC_m of each model (Busemeyer & Stout, 2002). The model with smaller BIC_m is preferred. The results show that $2*BIC_{ModelA} = 335.6$, and $2*BIC_{ModelB} = 316.8$. Hence Model B is selected to estimate the 6 threshold parameters and the parameter *s*.

Because the parameter distributions are skewed, the median is used to represent their central tendency. The median of *s* is 0.13. Medians of implicit threshold parameters are shown Figure 4. The implicit threshold curve is decreasing over turns.

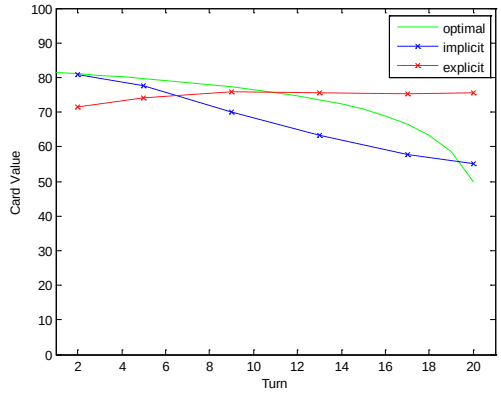

Figure 4: Explicit, implicit, and optimal thresholds.

Threshold Performances

The explicit, implicit, and optimal thresholds show considerable differences, with the first increasing over turns, while the implicit and optimal decrease. Moreover, in Figure 4, most parts of the implicit threshold are below the optimal. How do these differences in threshold values play out in terms of actual search performance? Does the explicitly stated threshold work better than the implicit threshold derived from subjects' actual choices, and how do both compare with the optimal strategy?

To answer these questions, first we linearly interpolated the explicit and implicit thresholds between the 6 known data points (turns 2, 5, 9, 13, 17, and 20) to obtain threshold values for all 19 turns (2-20), as shown in Figure 4. Then we performed 100,000 simulation runs for each of the three thresholds. The frequency distributions of performance (points per trial) for each are shown in Figure 5. All three distributions are negatively skewed. The frequency distribution of the implicit threshold is slightly more similar to that of the optimal strategy than is the explicit distribution, but all three are very similar. The mean and median of performance following the optimal strategy are 1601 and 1635; for the implicit threshold, 1595 and 1621; and for the explicit threshold, 1592 and 1603. The mean of subjects' actual performance on each trial, 1528, is a little farther away from the optimal performance, perhaps because of noise in subjects' choices or their use of different rules.

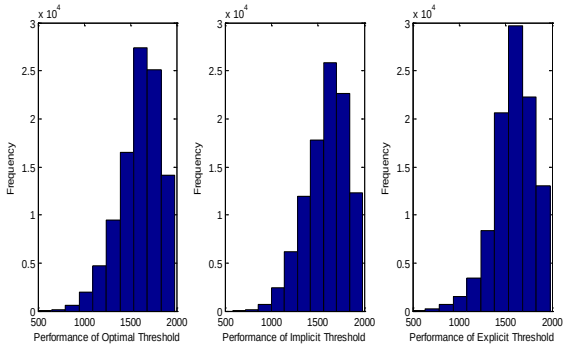

Figure 5: Frequency Distributions of Performance (Points Earned) by the Optimal, Implicit, and Explicit Thresholds.

Learning Effects

Although subjects do not know or follow exactly the optimal strategy, their implicit and explicit thresholds perform quite well—considering the noise in the actual data, these thresholds achieve impressively close to optimal results. How does this happen? Are subjects consistent in their performance across the 30 trials, or do they learn and improve based on the feedback provided after each trial?

To find out, we divided subjects' data into three parts according to trials. Data from the first trial to the 10th trial form the first part (F); the middle 10 trials are the second part (M); and the last 10 trials are the third part (L). Across

the three parts, we analyzed number of switches per trial, the turn number on which continuing exploitation commenced, and actual performance; these measures are shown in Table 1. Clearly, all three measures improved from the first to the last 10 trials, all coming closer to the optimal strategy. The frequency distributions of the starting turn of continuing exploitation of the three parts are shown in Figure 6 along with the optimal threshold's distribution. Again over trials, the distribution becomes more similar to the optimal one. Thus overall, learning occurs in terms of avoiding repeated switching between exploring and exploiting, and sticking to exploiting high-valued cards earlier, yielding increasing performance as well.

Table 1: Learning effects across trials from F to L.

	Number of switches	Starting turn of final exploitation	Performance
F	2.57	8.77	1492
M	1.54	6.9	1539
L	1.39	6.39	1553
Optimal	1	5.08	1601

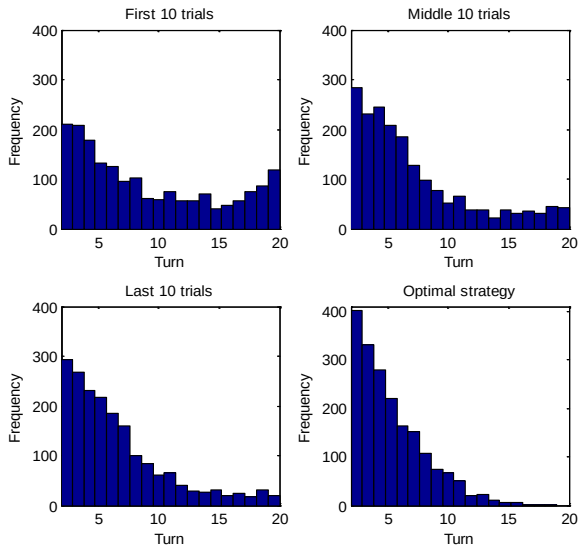

Figure 6: Distributions of turn number on which continuing exploitation commenced, changing over trials. Continued exploitation commences earlier as more trials are completed, and come close to the optimal distribution.

Finally, Model A was applied to each of the three parts of the data to estimate the implicit threshold across subjects for each of the ranges of trials. Because each subject only has 10 trials in each subset of the data, fewer data points can contribute to the modeling process, and if we tried to model implicit thresholds for subjects individually in these data subsets, there would be many subjects that both Model A and Model B would not fit well. To solve this problem, we combined all subjects' data of each subset together, and treated them as if they came from only one subject. *s* in Model B is a parameter of individuals' sensitivity. Given that all subjects' trials are combined together in each subset, the *s* in each section should be the same. In other words, it is no longer necessary to include this parameter *s*. Therefore instead of using Model B, Model A was selected to estimate the implicit threshold for each section.

The three implicit thresholds for the different trials are plotted in Figure 7, together with the optimal threshold. Basically, after turn 6, the implicit threshold value at each turn becomes smaller as the experiment continues (going from F to M to L). Overall, experience with the task leads subjects to more robustly use turn number as a factor in determining their thresholds.

For the first few turns, no matter which strategy someone uses, it is very important to set threshold values high enough to achieve a good performance. Consider the optimal strategy: the mean of the optimal threshold from turn 1 to turn 5 is around 80. If you explore consecutively for 5 times, the probability that you get at least one value higher than 80 among these 5 turns is about 70%. Most of time, this would let you achieve a good total score. But if you used a lower threshold value in the beginning, this would harm the final score substantially. Subjects also appear to learn this over multiple trials from F to M and L, increasing their implicit thresholds before turn 5.

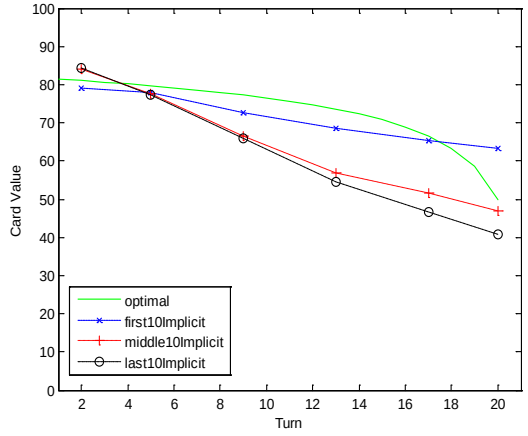

Figure 7: Implicit threshold curves found for the first, middle, and last trials across subjects (F, M, L), and the optimal threshold.

One interesting result is that after turn 5, implicit threshold curves diverge from the optimal one, and more strongly with more learning. Also surprisingly, the implicit thresholds go below 50 in M and L by the final turns (and it should never be appropriate to set a threshold for exploiting that is less than the mean value obtained from exploring, here 50). We think two possible reasons can account for these patterns. First, this could the result of noisy data

toward the end of each search trial. Given that subjects switch from exploring to exploiting at some point as turns go up, most of the data points at the end of turns involve exploiting a previously-found high value, rather than exploring the deck, so there is little data about when subjects would be willing to explore late in each trial. To maximize the log likelihood, Model B prefers to lower the corresponding threshold values as much as possible for those final turns, which would cause the implicit thresholds to become quite low (if also unreliable). The second reason is that subjects may really learn rules that direct them to decrease the implicit threshold for the last several turns. These explanations will be tested in further experiments.

Conclusions

The current paper addresses the issue of how people search an environment consisting of non-depleting resources by choosing between exploration and exploitation. The results indicate that subjects perform close to optimally, and get better over time with learning based on feedback. Subjects' mean total points per trial, number of switches between exploring and exploiting, and number of turns before starting continued exploitation become more similar to those of the optimal strategy as they go through more trials of searching. Subjects also appeared to adjust their implicit thresholds toward the optimal solution. The adjustment leads to a final implicit threshold that achieved a cumulative score quite close to the optimal one—even though that final implicit threshold has a simple linear shape, quite different from the accelerating falloff seen in the optimal threshold. It could be that the learning process is more adept at constructing a simple linear rule of this form than what optimal performance calls for; however, in this setting at least, performance hardly suffers as a consequence.

However, subjects themselves did not correctly report their use of a threshold that decreased over turns in each search trial: When asked to explicitly specify their thresholds, they stated ones that changed in the opposite direction of the implicit and optimal thresholds. This may have been due to subjects with little introspective insight just proposing that their threshold should increase as the trials increase, without thinking much more about the problem. In short, subjects do not explicitly know what is optimal nor what they are actually doing, as is often found in decision making tasks (Nisbett & Wilson, 1977), but they still get closer to optimal through a learning process.

There are several future directions that we are exploring. In the current project, the resources are non-depleting, and subjects have the ability to repeatedly shift between exploration and exploitation. But we can also use this setup to simulate depleting resources as in patch-based foraging and single choice searches with no recall as in the Secretary Problem, and investigate subjects' ability to learn appropriate strategies in those settings. We can also look for individual differences in tendency to explore versus exploit, and how that plays out across different search settings, including information search on the Web, as well as priming effects between settings. Finally, different populations may make the explore/exploit tradeoff in different ways, with some clinical populations emphasizing one aspect of search over the other (Hills, 2006); fMRI could also be useful in exploring these differences, as well as giving insights into the neural mechanisms used in search and whether they vary across different domains. By stripping search down to a setting where exploration and exploitation are most prominent, these comparisons may help us elucidate the underlying strategies more effectively.

Acknowledgments

We thank Ross Branscombe, Jerome R. Busemeyer, Woo-Young Ahn, and Thomas T. Hills for their help with this research. We also acknowledge the support of National Science Foundation REESE grant 0910218.

References

Beckage, N., Todd, P.M., Penke, L., and Asendorpf, J.B. (2009). Testing sequential patterns in human mate choice using speed dating. In Niels Taatgen and Hedderik van Rijn (Eds.), *Proceedings of the 2009 Cognitive Science Conference* (pp. 2365-2370).

Bell, W. J. (1991). *Searching behaviour: The behavioural ecology of finding resources.* New York: Chapman/Hall.

Busemeyer, J.R., & Stout, J.C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14,* 253-262.

Charnov, E.L. (1976). Optimal foraging: The marginal value theorem. *Theoretical Population Biology*, *9*, 129–136.

Ferguson, T.S. (1989). Who solved the secretary problem? *Statistical Science, 4(3),* 282–296.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*, 3–41.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555-580.

Livoreil, B., & Giraldeau, L.-A. (1997). Patch departure decisions by spice finches foraging singly or in groups. *Animal Behavior, 54,* 967–977.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Todd, P.M., & Miller, G.F. (1999). From pride and prejudice to persuasion: Satisficing in mate search. In G. Gigerenzer, P.M. Todd, and the ABC Research Group, *Simple heuristics that make us smart.* New York: Oxford University Press.

Wajnberg, E., Fauvegue, X., & Pons, O. (2000). Patch leaving decision rules and the marginal value theorem: An experimental analysis and a simulation model. *Behavioral Ecology, 11,* 577–586.