Final Report: Exploration and Exploitation in Structured Environments

Michael D. Lee and Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine

Abstract

This is the final report for the three-year AFOSR sponsored research project "Exploration and exploitation in structured environments" (FA9550-09-1-0082), with Michael Lee and Mark Steyvers as Co-PIs.

Executive Summary

In bandit problems, a decision-maker chooses repeatedly between a set of alternatives. They get feedback after every decision, either recording a reward or a failure. They also know that each alternative has some fixed unknown probability of providing a reward when it is chosen. The goal of the decision-maker is to obtain the maximum number of rewards over all the trials they complete.

Bandit problems provide an interesting formal setting for studying the balance between exploration and exploitation in decision-making. In early trials, it makes sense to explore different alternatives, searching for those with the highest reward rates. In later trials, it makes sense to exploit those alternatives known to be good, by choosing them repeatedly. How exactly this balance between exploration and exploitation should be managed, and should be influenced by factors such as the distribution of reward rates, the total number of trials, and so on, raises basic questions about adaptation, planning, and learning in intelligent systems.

This research project completed a series of inter-related lines of bandit problem research that improved our understanding of human and optimal sequential decision-making using bandit problems, covering the following topics:

Heuristic models. We have developed a new heuristic model of human decision-making, relying on the idea that people switch between latent states of exploration and exploitation. This model performs well in accounting for both optimal and human performance, when compared to standard heuristics from the reinforcement learning literature. We also show how inferring the psychologically meaningful parameters for the new heuristic provides a simple and interpretable account of optimal decision-making, human decision-making, and the relationship between the two.

Individual differences. In a range of different bandit problem experiments, we have observed a significant range of individual differences in decision-making. In one large study with 451 participants, we also collected various measures of cognitive abilities and personality variables, and found some interesting correlations between these characteristics, and how people managed to exploration vs exploitation trade-off in solving bandit problems. We also considered a non-parametric Bayesian modeling approach to the individual differences.

Contaminant processes. Individual differences raise the challenge of filtering out participants who used overly-simple and uninteresting cognitive processes. We report on a latent mixture model that identifies these people, using a wide range of candidate models of contaminant behavior, and show how their removal affects parameter inference for the substantive decision-making models.

Adaptation to change. We studied how people learn and adapt in bandit problems where the environment changes, and report on particle filtering models of this behavior.

Wisdom of crowds. By aggregating over the behavior of many participants to infer model parameters, we explored the possibility of a "wisdom of the crowds" effect for bandit problems, whereby group behavior outperforms all or the majority of individuals.

Design optimization. We applied design optimization methods from statistics to the problem of creating optimal bandit problems to distinguish competing models, and report on the insights provided by the application of these methods.

In this report, we summarize the main research achievements and highlights, giving references to the published papers for each topic that provide full details.

Heuristic Models

- Lee, M.D., Zhang, S.,Munro. M.N., & Steyvers, M. (2009). Using heuristic models to understand human and optimal decision-making on bandit problems. In *Proceedings of the Ninth International Conference on Cognitive Modeling.*
- Lee, M.D., Zhang, S., Munro, M.N., & Steyvers, M. (accepted, pending minor revisions). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research.*
- Zhang, S., Lee, M.D., & Munro. M.N. (2009). Human and optimal exploration and exploitation in bandit problems. In *Proceedings of the Ninth International Conference on Cognitive Modeling.*

Lee, Zhang, Munro, and Steyvers (in press) provide a consolidated overview on work developing and comparing a set of heuristic models of people's decision-making on bandit problems. The work developed a new heuristic, called τ-switch, based on latent switching between exploration and exploitation, and compared it with the benchmark win-stay lose-shirt, ϵ-greedy, ϵ-decreasing and ϵ-first algorithms from the reinforcement learning literature (e.g., Sutton & Barto, 1998).

The key to deriving the new τ-switch heuristic came from a modeling analysis involves the pattern of change between latent exploration and exploitation states in a very general latent-state model. This analysis is reported in detail by Zhang, Lee, and Munro (2009), and is summarized in Figure 1. This figure shows whether the model is an exploration or exploitation state as it accounts for both the human and optimal data, over six experimental conditions. The experimental conditions are organized into the panels, with rows corresponding plentiful, neutral and scarce environments, and the columns corresponding to the 8- and 16-trial problems. Each bar graph shows the probability of an exploitation state for each trial, beginning at the third trial (since it is not possible to encounter the explore-exploit situation until at least two choices have been made). The larger bar graph, with darker blue bars, in each panel is for the optimal decision-making data. The 10 smaller bar graphs, with lighter green bars, corresponds to the 10 subjects within that condition.

The most striking feature of the pattern of results in Figure 1 is that, to a good approximation, once the optimal or human decision-maker first switches from exploration to exploitation, they do not switch back. There are some exceptions—both participants RW and BM, for example, sometimes switch from exploitation back to exploration briefly, before returning to exploitation—but, overall, there is remarkable consistency. Most participants, in most conditions, begin with complete exploration, and transition at a single trial to complete exploitation, which they maintain for all of the subsequent trials. This general finding remarkable, given the completely unconstrained nature of the model in terms of exploration and exploitation states. All possible sequences of these states over trials are given equal prior probability, and all could be inferred if the decision data warranted.

Figure 2 examines the ability of the τ-switch model coming from this analysis in account for human decision-making, relative to the machine learning benchmarks, and shows the posterior predictive average agreement of each model to individual participant. Participants are shown as bars against each of the models. We conduct analysis at the level of individual participants to allow for the possibility of individual differences. This intuition seems to be borne out. For the first 8 of the 10 participants (shown in darker blue), the τ-switch models provides the greatest level of agreement. For the last 2 of the 10 participants (shown in lighter yellow), this result is not observed, but it is clear that none of the models is able to model these participants well. One possibility is that these participants may have changed decision-making strategies during completing the 50 problems, and this prevents any single model from providing a good account of their performance.

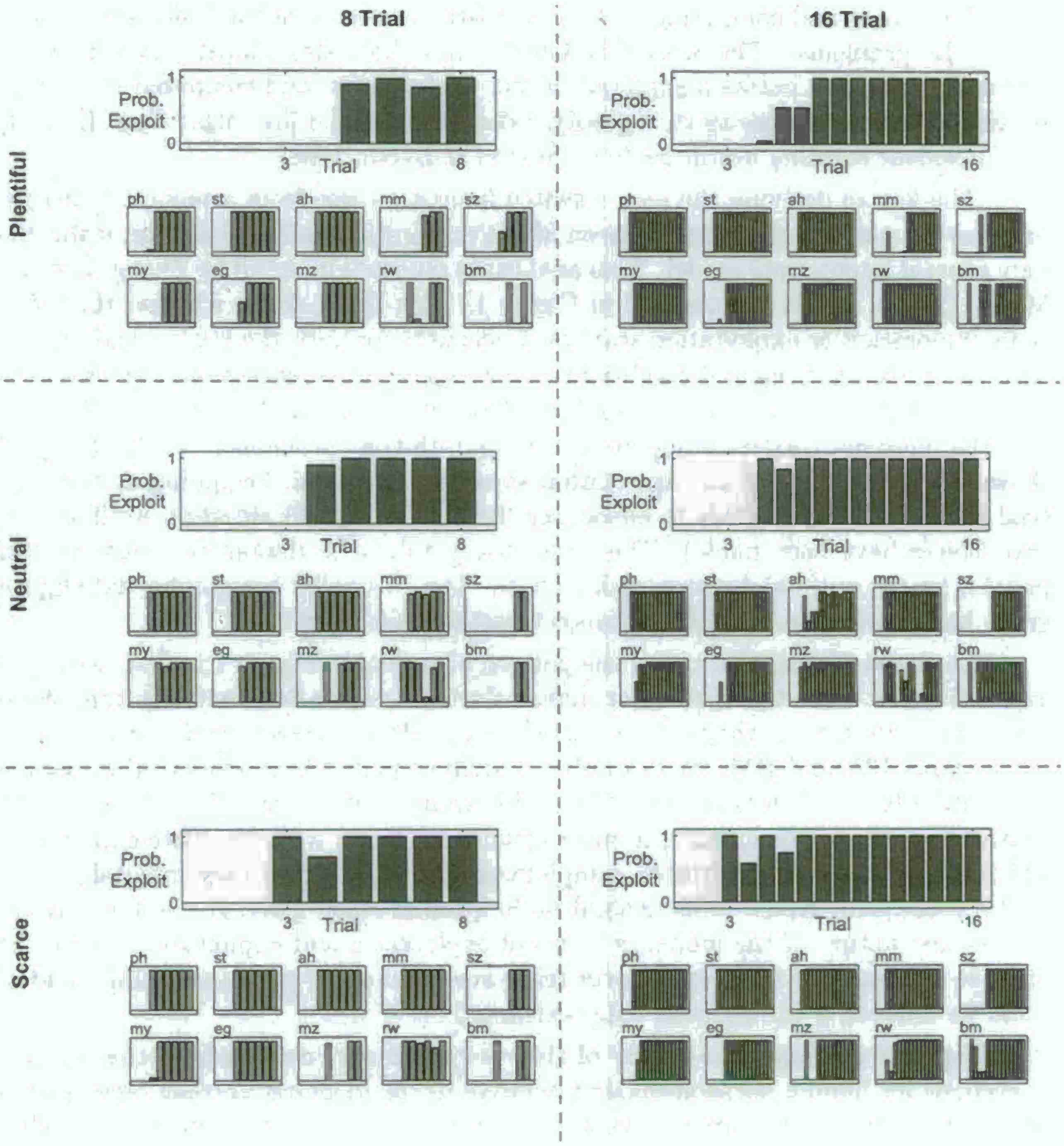

Figure 1. Each bar graph shows the inferred probabilities of the exploitation state over the trials in a bandit problem. Each of the six panels corresponds to an experimental condition, varying in terms of the plentiful, neutral or scarce environment, or the use of 8 or 16 trials. Within each panel, the large blue (darker) bar graph shows the exploitation probability for the optimal decision-process, while the 10 smaller green (lighter) bar graphs correspond to the 10 participants.

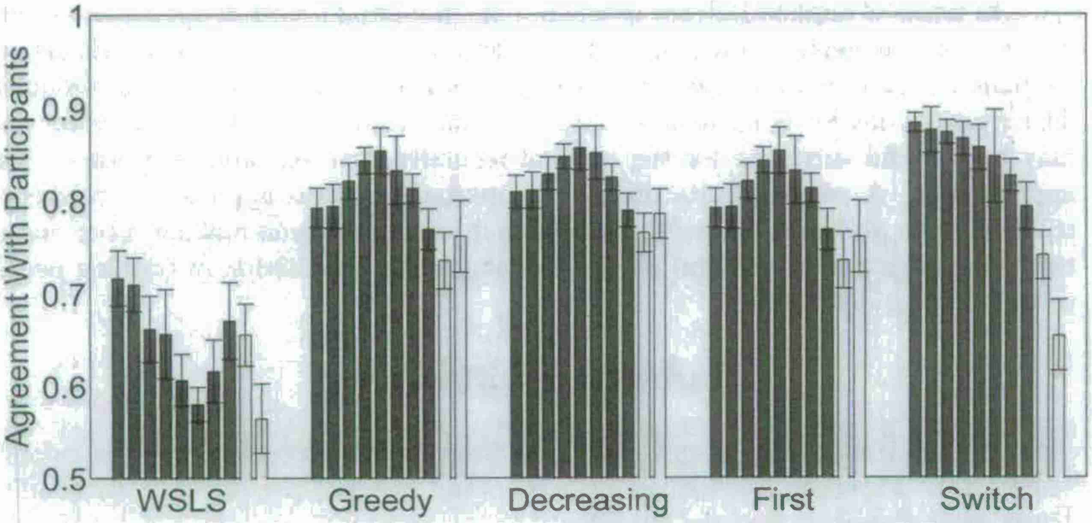

Figure 2. Posterior predictive average agreement of the heuristic models with each individual participant. Two 'outlier' participants, not modeled well by any of the heuristics, are highlighted in lighter yellow.

Overall, however, our results show that, for the large majority of participants well described by any model, the τ-switch model is the best. In fact, Figure 2 suggests that the ability of the model to model human decision-making follows the same ordering as their ability to mimic optimal decision-making. WSLS is the worst, followed by the three reinforcement learning models, which are approximately the same, and then slightly improved by the new τ-first model.

The key overall finding from this project is that the τ-switch model is a useful addition to current models of finite-horizon two-arm bandit problem decision-making. Across the three environments and two trial sizes we studied, it consistently proved better able to mimic optimal decision-making than classic rivals from the statistics and machine learning literatures. It also provided a good account of human decision-making, for the majority of the participants in our study. To this end, the model comparisons we have done have theoretical implications for understanding the nature and limitations of human decision-making. Our work also illustrated a useful general approach to studying decision-making using simple heuristic cognitive models. Three basic challenges in studying any real-world decision-making problem are to characterize how people solve the problem, characterize the optimal approach to solving the problem, and then characterize the relationship between the human and optimal approach. Our results show how the use of simple heuristic models, using psychologically interpretable decision processes, and based on psychologically interpretable parameters, can aid in all three of these challenges.

In terms of applied Defense outcomes, one potential practical application of our new τ-switch model is to any real-world problem where a short series of decisions have to made be made with limited feedback, and with limited computational resources. The τ-switch model is extremely simple to implement and fast to compute, and may be a useful surrogate for the optimal recursive decision process in some niche applications. A second, quite different, potential practical application, relates to training. The ability to interpret optimal and human decision-making using one or two psychologically meaningful parameters could help instruction in training people to make better decisions.

Individual Differences

- Lee, M.D., Zhang, S., Munro, M.N., & Steyvers, M. (accepted, pending minor revisions). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research.*
- Steyvers, M., Lee, M.D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53*, 168–179.
- Zeigenfuse, M.D., & Lee, M.D. (2009). Bayesian nonparametric modeling of individual differences: A case study using decision-making on bandit problems. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Shonmaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 1412-1415. Austin, TX: Cognitive Science Society.

Steyvers, Lee, and Wagenmakers (2009) applied four models to data from 451 human participants, using the Bayes Factor to choose which model provided the best account of each individual. The four models were a simple guessing model, a version of the class win-stay lose-shift model, a model based on the observed success rate of each alternative, and the optimal model calculated using standard dynamic programming methods (e.g., Kaebling, Littman, & Moore, 1996).

The results are summarized in Figure 3. The left panel shows the distribution of the log Bayes Factor measure. The right panel shows the break-down of the participants into the proportions who best supported each of the four models. There is clear evidence of individual differences in Figure 3, with a significant proportion of participants being most consistent with all three of the optimal, success ratio, and win-stay lose-shift models. Interestingly, about half of our participants were most consistent with the psychologically simple win-stay lose-shift strategy, while the remainder were fairly evenly divided between the more sophisticated success ratio and optimal models. Very few participants provided evidence for the guessing model, consistent with these participants being 'contaminants', who did not try to do the task.

One interpretation of these results is that subsets of participants use successively

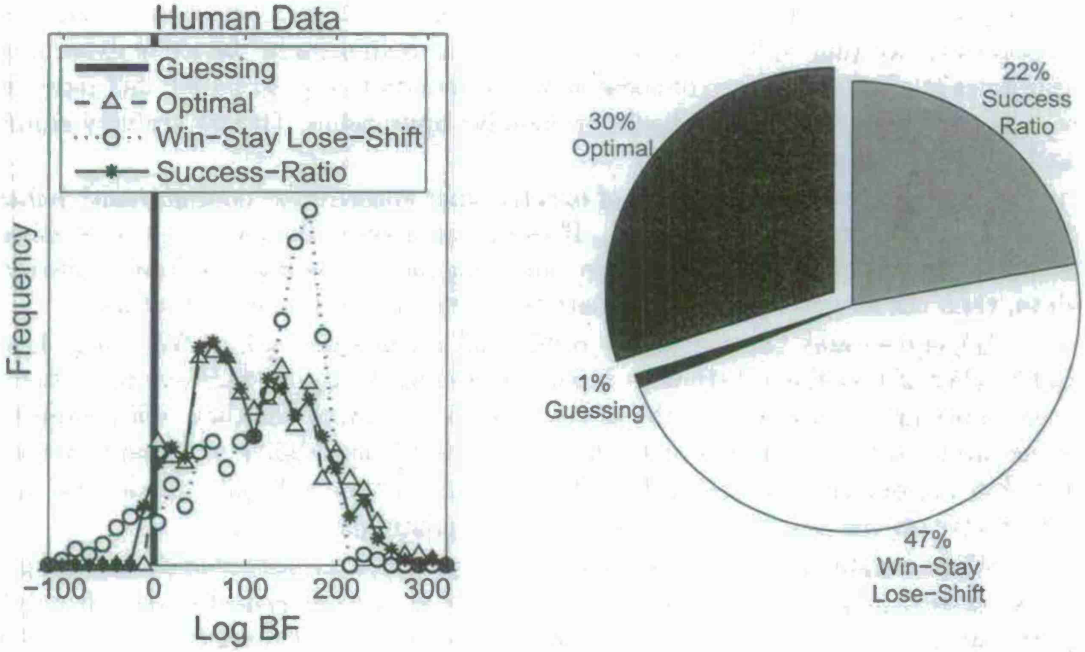

Figure 3. The distribution of the log BF (log Bayes Factor) measures over the participant data (left panel), and the sub-division into the proportions best supported by each of the four models (right panel).

more sophisticated decision-making strategies. The win-stay lose-shift decision model does not involve any form of memory, but simple reacts to the presence or absence of reward on the previous trial. The success ratio model involves comparing the entire reward history of each alternative over the course of the game, and so does require memory, but is not explicitly sensitive to the finite horizon of the bandit problem. The optimal model is sensitive to the finite horizon, and to the entire reward history, and so involves trading off exploration and exploitation. Figure 3 suggests that sizeable subsets of participants fell at each of these three levels of psychological sophistication.

Contaminant Processes

- Zeigenfuse, M.D., & Lee, M.D. (in press). A general latent-assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*.

One interesting issue in studying human performance on bandit problems involves the potential use of different decision-making strategies. There are many psychologically plausible heuristic approaches coming from the game theory and reinforcement learning literatures (e.g. Sutton & Barto, 1998), as well as heuristics developed in the cognitive sciences (e.g. Zhang et al., 2009), and there is some empirical evidence that different people use different heuristics in the same experiment (Steyvers et al., 2009). Some of these heuristics are quite sophisticated, and represent what might be viewed as intelligent or effective approaches. Others are very simple, and clearly sub-optimal.

This contrast raises the issue of exactly what constitutes "contaminant" behavior in a bandit problem experiment. If the focus is on understanding the relatively sophisticated models by, for example, inferring model parameters from behavioral data, then the simple heuristic approaches can be viewed as contaminating.

Zeigenfuse and Lee (in press) conducted an analysis using "Win-Stay Lose-Shift" (WSLS) as the substantive model (Robbins, 1952). WSLS assumes that if, after choosing an alternative, the decision-maker is rewarded, they will choose the same alternative on the next trial with some (high) probability γ. Alternatively, if the decision-maker is not rewarded, WSLS assumes they will only choose the same alternative on the next trial with some (small) probability $1 - \gamma$.

While extremely simple, the WSLS often provides a reasonable account of people's decision-making. For example, Steyvers et al. (2009) collected data from 451 participants on a series of bandit problems, and presented a series of model comparisons showing that the majority of these participants decisions consistent with WSLS. We use an abbreviated version of the same data set—using a subset of participants chosen to make clear the contaminant modeling principles this example aims to explain—including 47 participants. As with the full data set, all participants completed a set of 20 bandit problems, each involving four alternatives and 15 trials.

Zeigenfuse and Lee (in press) also considered two plausible strategies a non-motivated participant might use to complete the task. One, called the 'random' strategy, involved simply chosing an alternative at random on every trial. The other non-motivated strategy was called 'same', and involved the participant choosing the same alternative on almost every trial, regardless of the observed pattern of reward.

Zeigenfuse and Lee (in press) applied the three models—the substantive WSLS, and the contaminant random and same heuristics—to the Steyvers et al. (2009) data in four separate analyses, all using Bayesian latent mixture modeling to identify contaminants. In the first, they simply applied the WSLS model. In the second analysis, they applied WSLS, but also introduced the random model as a contaminant model, using the latent assignment approach. In the third analysis they applied WSLS with the same model as the contaminant model. In the fourth analysis, they used both the random and same models as contaminants, allowing the behavior of each participant to be explained by any one of these three accounts.

The left panel of Figure 4 shows how the participants were assigned to the three models. Each point corresponds to a participant, and the type of marker indicates whether they were classified as following the WSLS, random or same model. The axes in which the points are displayed correspond to two summary measures of their decision-making, chosen because they capture much of the variance involved in par-

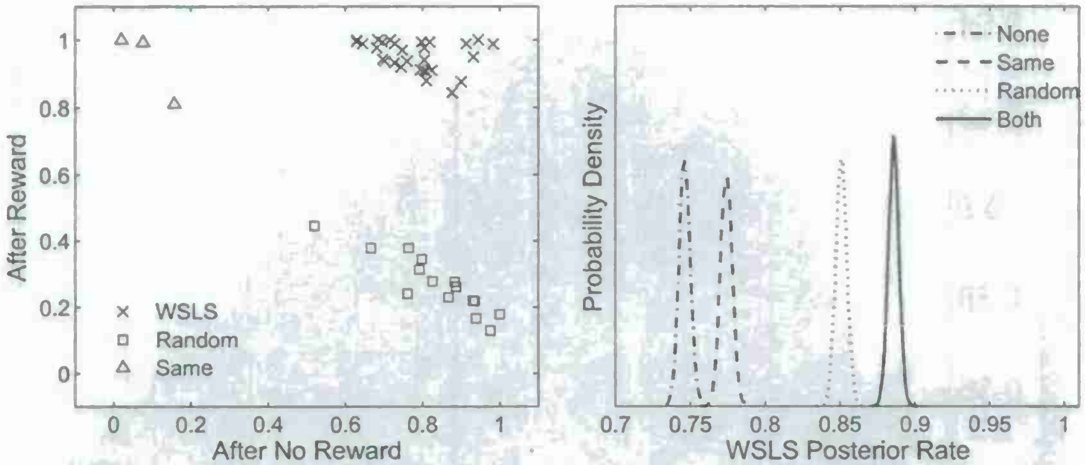

Figure 4. Analysis of bandit problem behavior. The left panel shows the 47 participants, and their assignment to the WSLS, random and same models. The right panel shows the posterior distribution of the WSLS rate γ for four analyses, including no contaminant modeling, the random contaminant model, the same contaminant model, or both contaminant models.

titioning the participants among the models. The x-axis shows the proportion of trials following no reward that a different alternative was chosen on the next trial. The y-axis shows the proportion of trials following a reward the same alternative was chosen on the next trial.

WSLS performance corresponds to high values on both measures, and so these participants are in the top-right corner. Random model performance corresponds to the point $(0.75, 0.25)$, since there are four alternatives. Same model performance correspond to the top-left corner of the graph. The left panel of Figure 4 shows a clear partitioning of participants into each of these regions, and that they are appropriately assigned by the model. In other words, there are clear individual differences between participants in the decision strategy these use to solve bandit problems, and they appear to be well described by the WSLS, random and same models for these participants.

The inferences about the γ parameter of WSLS are shown, for all four analyses, in the right panel of Figure 4. The key point is that the inferred rate of winning and staying or losing and shifting changes significantly depending on the assumptions made about contaminant behavior in the participant pool. When no contamination is assumed, γ is around 0.75. When both the same and random forms of contamination are included in the analysis, the inferred γ increases to almost 0.9. Using just one or other of the contaminant models gives different intermediate values. These results make clear that what is learned by applying a substantive cognitive model to behav-

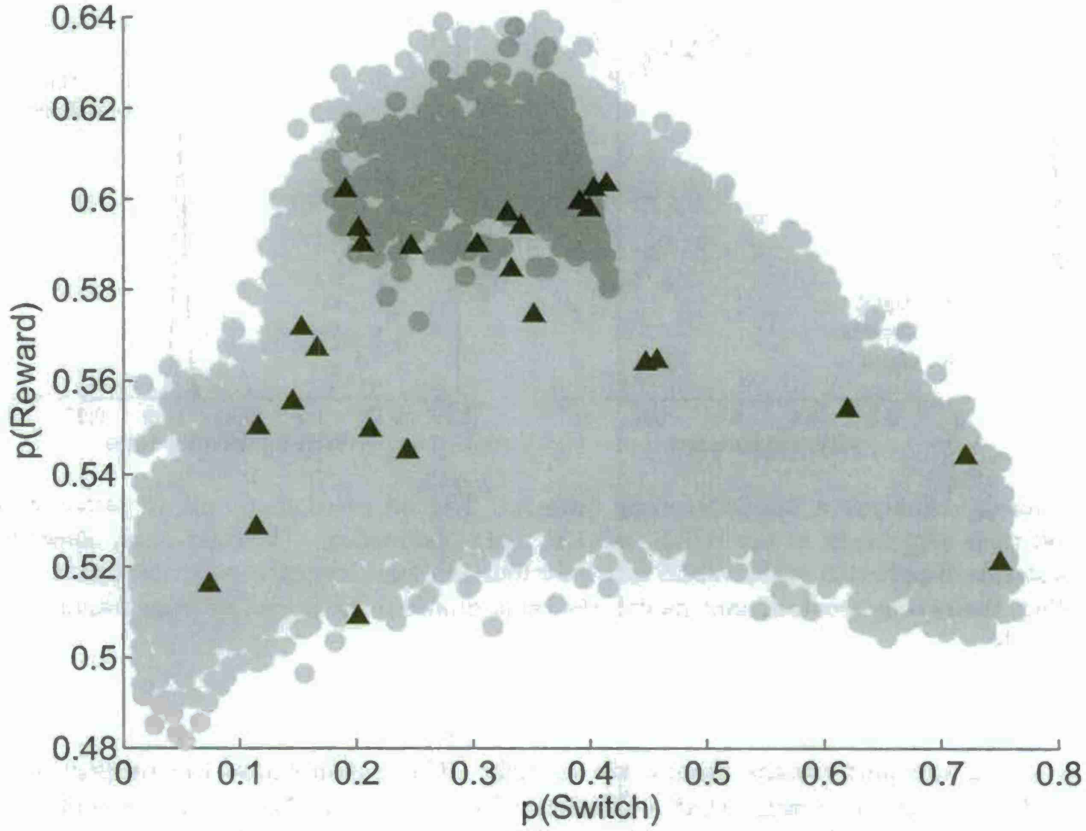

Figure 5. Subject performances in (black triangles) against the range of the continual-change reward rate particle filter (green or darker gray) and discrete-change reward rate particle filter (lighter gray).

ioral data can depend critically on the nature of possible contamination processes included in the analysis.

Adaptation to Change

- Yi, S.K.M., Steyvers, M., & Lee, M.D. (2009). Modeling human performance in restless bandits using particle filters. *Journal of Problem Solving, 2,* 33-53.

Yi, Steyvers, and Lee (2009) investigated 'restless' bandit problems, where the distributions of reward rates for the alternatives change over time. This dynamic environment encourages the decision-maker to cycle between states of exploration and exploitation. In one environment we consider, the changes occured at discrete, but hidden, time points. In a second environment, changes occured gradually across time. Decision data were collected from people in each environment. Individuals varied substantially in overall performance and the degree to which they switched between alternatives.

Yi et al. (2009) modeled human performance in the restless bandit tasks with two particle filter models, one that can approximate the optimal solution to a discrete restless bandit problem, and another simpler particle filter that is more psychologically plausible. The key result was that the simple particle filter was able to account for most of the individual differences. This result is summarized in Figure 5, which shows the range of human performance (black triangles) against the range of the optimal model (green or dark gray), and the psychologically-plausible sub-optimal model (lighter gray). The sub-optimal model propagates particles depending on just a simple estimated reward rate (i.e., a cognitively plausible summary of the full uncertainty about reward rates). It is clear that the additional variation in performance of this sub-optimal model is required to describe the variation in behavior seen in people.

Wisdom of Crowds

- Zhang, S., & Lee, M.D. (in press). Cognitive models and the wisdom of crowds: A case study using the bandit problem. In R. Catrambone, & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

An enticing idea in the study of individual and group decision-making is the phenomenon known as the "wisdom of crowds". The idea is that, by aggregating the behavior of a group of people doing a challenging task, it is possible for group performance to match or exceed the performance of any of the individuals. Surowiecki (2004) provides an extensive survey of wisdom of crowds results over a diverse set of human endeavors and decision-making situations, ranging from guessing the weight of an ox at a county fair, to inferring the location of a missing submarine, to predicting the outcome of sporting events. While the exact conditions needed for group performance to exceed individual performance are not completely understood, it seems clear that crowds can be wise in any situation where people have some partial knowledge, and the gaps in their knowledge are subject to individual differences. Under these circumstances, aggregation of individual decisions can serve to amplify the common signal and reduce the idiosyncratic noise, leading to superior group performance.

One challenge in producing wisdom of crowds effects arises when tasks are more complicated than estimating a single quantity, or predicting a simple outcome. Many interesting and real-world decision-making situations are inherently multidimensional or sequential. In these situations, it is often not possible to combine the raw behaviors of people, because they are not commensurate. For example, imagine trying to combine the expertise of basketball fans trying to predict the result of an eight-team single elimination tournament, with quarter-finals, semi-finals and a final. Based on their decisions about the quarter-finals, these people may be making decisions about different teams in the semi-finals and final. This makes simple aggregation based on their raw decisions impossible for the later rounds.

For more difficult decision problems like these, we believe cognitive science has a key role to play in wisdom of the crowd research. Rather than aggregating people's behaviors, it is necessary to aggregate their knowledge, as *inferred* from their behavior. This inference needs models of cognition, accounting for how latent knowledge manifests itself as observed behavior within the constraints of a complicated task.

Zhang and Lee (2010) completed a case study of the application of cognitive models for bandit problems. By applying a series of existing models of human decision-making on the task to a variety of data sets, they showed that it is possible to produce aggregate performance that is near optimal, and far exceeds the performance of most of the individuals. The analysis involved taking a set of standard decision-making models, and using the inferred group mean in a hierarchical Bayesian analysis. This gives a natural model-based aggregation of individual performance, and solves the problem of aggregating the knowledge of different people solving different, but related, bandit problems. Rather than aggregating their behavioral choices, we are aggregating the psychology parameter values that lead to those choices. To complete the model-based wisdom of crowd analyses, we used the group mean parameter values to define a "group model" that used the same decision-process, and completed the same problems given to participants in each of the three experiments. Because the number of rewards obtained is inherently stochastic, we repeated this many times to approximate the distribution of rewards. We also applied the optimal decision-making process to each experiment, to approximate the best possible distribution of rewards for each experiment.

The results are shown in Figure 6. The columns correspond to the three experiments. The rows correspond to the WSLS, extended WSLS, ϵ-greedy and ϵ-decreasing decision models. Within each panel, the squares piled into histograms show the distribution of performance (i.e., how many rewards were obtained) for the individual participants. The two curves then correspond to the distribution of performance for the group model (red, dotted line) and the optimal decision process (green, solid line).

Figure 6 shows that some of our decision-making models do produce a clear wisdom of the crowds effect, whereas others do not. The distributions of rewards for the group model formed by the WSLS and extended WSLS models does not improve on the distribution of individual performance, and are not close to optimal. For the ϵ-greedy and ϵ-decreasing group models, however, there is significant improvement. In particular, the ϵ-decreasing group model has a distribution of rewards that is extremely close to the optimal distribution for all three experiments.

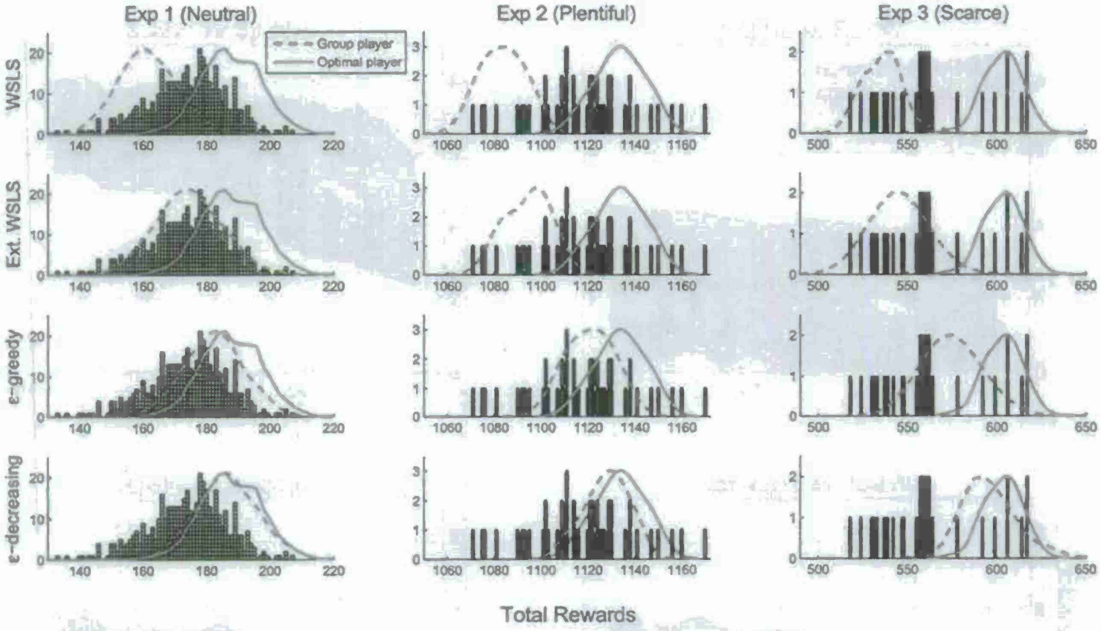

Figure 6. Distribution of rewards for individual participants, the group model, and the optimal decision-making process, for each decision-making model and each experiment.

Design Optimization

- Zhang, S., & Lee, M.D. (submitted). Optimal experimental design for a class of bandit problems. Submitted to the *Journal of Mathematical Psychology*.

A basic challenge for measuring human performance in bandit problem—a key part of our project—is to design experiments that will provide the most useful data. Traditionally, psychological experiments have been designed to meet these goals based on a mixture of previous results, pilot information, and the intuition of the experimenter. This is the approach we originally took in Steyvers et al. (2009). Formal approaches to experimental design optimization, however, have received considerable attention in statistics and engineering, and, recently, psychologists have also started to search for approaches that allow the formal optimization of the design of an experiment (e.g. Myung & Pitt, 2009).

In Zhang and Lee (submitted), we adapted the formal framework for experimental design optimization described by Myung and Pitt (2009) to a research area where it has not previously been applied. We developed MCMC algorithms for design optimization, tailored to answer the question of how bandit problem experiments with people should be designed, so as to maximize the usefulness of the data in dis-

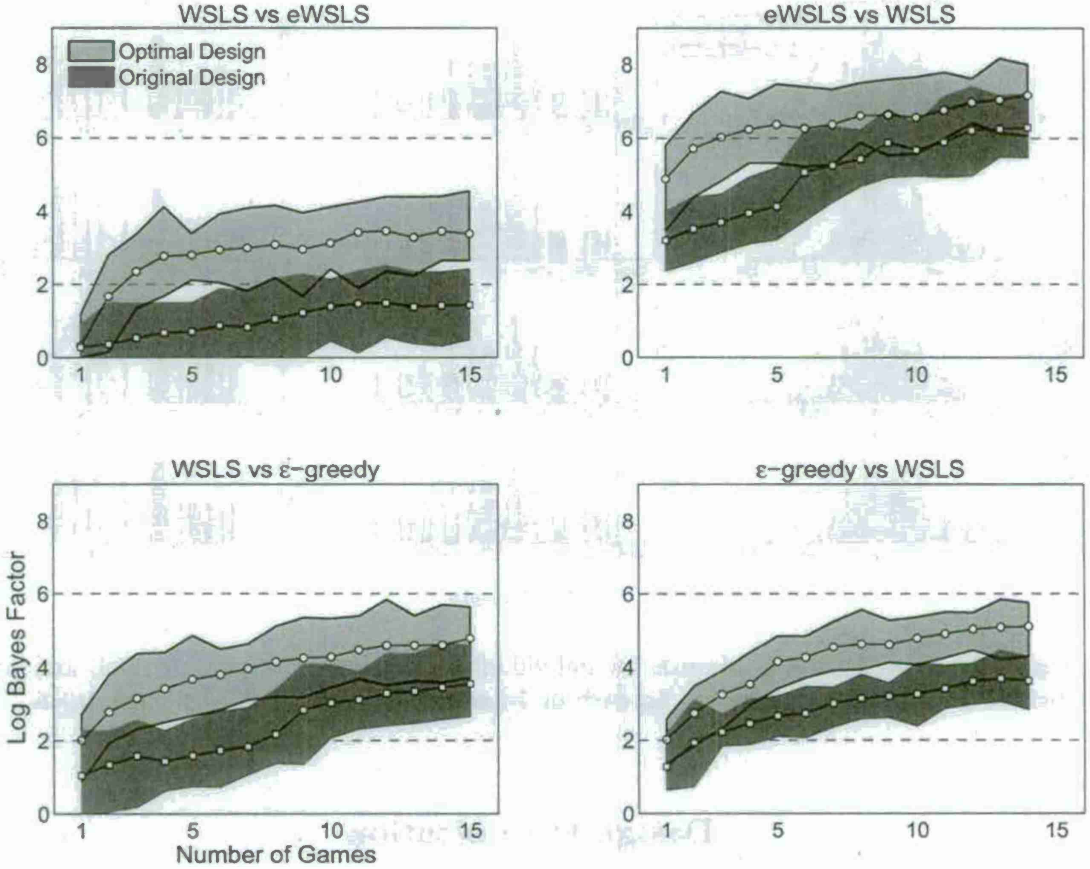

Figure 7. Performance of optimal experiments relative to the original design used by Steyvers et al. (2009).

tinguishing competing models of human cognition.

Figure 7 shows the final results of this work. In these analyses, one of the candidate models is used to generate decision-making data for a sequence of bandit problems following either the optimal design, or the original design used by Steyvers et al. (2009). Under both designs, the log Bayes Factor in favor of the correct generating model is used to measure the effectiveness of the experimental design. Figure 7 shows the mean (by lines and markers) and the range (by bounded shaded regions) for the log Bayes Factors, in four different analyses. These consider both the WSLS vs eWSLS and WSLS vs ϵ-greedy model comparisons, and consider both assumptions about which model generated the data. The means, minima and maxima shown are based on 100 independent runs of each simulated experiment. It is clear from Figure 7 that the optimal design always outperforms the original design on average.

Even more compellingly, the worst observed optimal design is always better than the mean original design, and is often better than the best-performed original design.

References

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Lee, M. D., Zhang, S., Munro, M. N., & Steyvers, M. (in press). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research*.

Myung, J., & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *55*, 527–535.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Random House.

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

Yi, S. K. M., Steyvers, M., & Lee, M. D. (2009). Modeling human performance in restless bandits using particle filters. *Journal of Problem Solving*, *2*, 33–53.

Zeigenfuse, M. D., & Lee, M. D. (in press). A general latent-assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*.

Zhang, S., & Lee, M. D. (2010). Cognitive models and the wisdom of crowds: A case study using the bandit problem. In R. Catrambone & S. Ohlsson (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Zhang, S., & Lee, M. D. (submitted). Optimal experimental design for a class of bandit problems.

Zhang, S., Lee, M. D., & Munro, M. N. (2009). Human and optimal exploration and exploitation in bandit problem. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the Ninth International Conference on Cognitive Modeling — ICCM2009*. Manchester, UK.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
28-06-2010

2. REPORT TYPE
Final Report

3. DATES COVERED *(From - To)*
01-Jan-2007 - 31-Mar-2010

4. TITLE AND SUBTITLE
MODELING EXPLORATION AND EXPLOITATION IN STRUCTURED ENVIRONMENTS

5a. CONTRACT NUMBER
FA9550-07-1-0082

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Lee, M. D. and Steyvers, M.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California, Irvine
Irvine CA, 92967

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 N Randolph St
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0078

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This is the final report for the three-year AFOSR sponsored research project ``Exploration and exploitation in structured environments" (FA9550-09-1-0082), with Michael Lee and Mark Steyvers as Co-PIs.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
U	U	U	UU	15	19b. TELEPHONE NUMBER *(Include area code)*

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.