Towards a Theory of Balancing Exploration and Exploitation in Probabilistic Environments

Stefani Nellen (snellen@andrew.cmu.edu)
Marsha C. Lovett (lovett@cmu.edu)
Carnegie Mellon University,
Department of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15213, USA

Abstract

Learning to make good choices in a probabilistic environment requires that the Decision Maker resolves the tension between exploration (learning about all available options) and exploitation (consistently choosing the best option in order to maximize rewards). We present a mathematical learning model that makes selections in a repeated-choice probabilistic task based on the expected payoff associated with each option *and* the *information gain* that will result from choosing that option. This model can be used to analyze the relative impact of exploration and exploitation over time and under different conditions. It predicts the aggregated and individual learning trajectories of participants in various versions of the task sufficiently well to support our basic argument: *Information gain* is a valid and rational criterion underlying human decision making. Future modeling work will be addressing the exact nature of the interaction between exploration and exploitation.

Introduction

Decision makers are often placed in novel situations that offer them a finite variety of choices. They know that each of the choices is associated with some probability of leading to a positive outcome, but they don't know what these probabilities are. They might also know that making a choice will constrain the available choices in the subsequent decision cycle, but, again, they don't quite know in which manner this will happen. All they know is that they will have *n* opportunities to make one choice at a time, and that the long-term goal is to maximize accumulated rewards. Each Decision Maker has to resolve the tension between exploration (learning about the payoffs of the options, which is achieved by selecting them and observing the outcome) and exploitation (consistently choosing the best option). Maximizing rewards depends on accurate payoff estimates and therefore on sufficient information. On the other hand, this exploration must eventually be discarded in favor of consistently choosing the best option if the goal is to be met. This example shows that exploration is a necessary prerequisite of probability learning. It also shows that the benefit of exploration depends on the amount of information that has already been accumulated. Therefore, it will not remain constant over time. We are convinced that it is possible to predict the relative impact of exploration and exploitation under different conditions over time, and that this relative impact varies in interaction with the probabilistic structure provided in the environment. The model we present here provides an implementation of this basic idea.

In the context of decision making research, the representation of probabilistic information/learning has been somewhat understudied, partly because probabilistic representations are often assessed by eliciting one-shot probability estimates from participants instead of observing changes in their actual behavior (see Gigerenzer, 1994, for a critique of the notion of single-event probabilities). Additionally. following a long history of models of probability learning (e.g. Estes, 1964;), recent cognitive models of heuristics or decision making algorithms that rely on probabilistic cues often provide participants with the explicit probabilities from the outset (Broeder, 2000; Payne, Bettman and Johnson, 1993; Rieskamp & Hoffrage, 2000) or assume that the representations have already been formed (models often achieve this during a separate "training phase") and are ready to be used. In ACT-R (Anderson & Lebiere, 1998), the selection of cognitive operators (production rules) is based on their "expected utility", which is partly determined by estimates of the probability of success associated with that operator (the other components being "cost/ effort" and noise). This expected utility can gradually be learned by experience. *Information gain*, i.e. the increase in knowledge about the payoff-structure of the system, does not play an explicit role in utility learning and operator selection in any of these models right now.

The model presented in this report proposes a definition of *information gain* and an explanation of information gain and estimated payoffs interactively influence decision making. The model also explains how the relative impact of information gain and estimated payoffs on decision making changes over time.

Before describing the model and its predictions in more detail, we will provide a description of the task in which it was developed and tested, a repeated-choice probability learning task with immediate and delayed rewards. Afterwards, we will present the fit between some model predictions and behavioral data, compare this fit to that of a version of the model that does not take *information gain* into account, and explore future directions for refining the model.

A Probabilistic Learning Task

A version of the task that was used for developing this model has previously been used by Brown and Lovett (2001) in order to assess people's ability to learn to prefer long-term over short-term benefits. In that context, it has been dubbed a "Single Player Version of the Prisoner's Dilemma (PD)". However, the particular connotations of the PD-game are of less relevance for the present work and might be confusing here. For this reason, we will simple be referring to the "Probabilistic Learning Task (PLT)".

Figure 1 shows a schematic overview of the PLT, including its underlying rewards and constraints structure. Participants go through the task at the computer. They are presented with four options in the form of four closed doors, two red ones and two green ones (in the figure, the dark doors in the top row are both red). In the instructions, participants are given the following pieces of knowledge about the task ahead of them: They know that they are asked to make a series of choices, selecting one door at a time. They need a key to open a door of the same color, and they are given a red key at the beginning of the task. By opening a door, the current key will be given up. Upon opening a door, two things will happen: The participants may or may not receive a reward (5 "points"). Additionally, they will receive a new key that will be either red or green. This key will constrain their available options in the subsequent trial, because red keys open only red doors and green keys open only green doors. Finally, participants are told that the goal in this game is to gain as many points as possible.

Upper Left (UL):	Upper Right (UR):
p(reward)=0.6 Key: Red Door Color: red	p(reward)=0.8 Key: Green Door Color: Red
Bottom Left (BL):	Bottom Right (BR):
p(reward)=0.2 Key: Red Door Color: Green	p(reward)=0.4 Key: Green Door Color: Green

Figure 1: Overview of the choices and their outcomes in the PLT. The outcomes are shown for clarification. Participants have to learn them .

It may be noted that green doors have a lower probability of reward than red doors. However, the upper right red door (UR), which has a higher reward probability than the upper left red door (UL) gives a green key. To solve the game, participants have to learn that UL has the highest payoff probability in the long run, even though its immediate rewards are less probable than those of UR. The behavioral manifestation of having solved the game is to choose UL consistently, disregarding the other options.

Before being able to solve the game, participants must learn the reward-probabilities associated with each door and the association between door and keys (the latter connection is deterministic). Participants complete a total of 200 trials in one session with this task.

The "Expected Utility Differences (EUDs)" in Two Versions of the Task

Given the constraint of the keys, there are three sequences of choices that can be done repeatedly: Choosing UL a number of times, choosing first UR and then BL, and Choosing BR repeatedly. These three strategies can be evaluated in terms of the payoffs associated with the involved doors. In the PLT, the expected values of the strategies follow the following general constraint:

$$2*payoff(UL) > payoff(UR)+payoff(BL) > 2*payoff(BR)$$

The payoff probabilities of the doors can be manipulated to characterize different versions of the PLT, which differ in the extent of the "greater than" relation. This relation is called the "Expected Utility Difference" (EUD). Inserting the reward probabilities from fig. 1 in the equation above and multiplying the outcome with five (because of the "points" that form the reward) yields

$$2*(0.6*5) > (0.8*5+0.2*5) > 2*(0.4*5), \text{ or}$$
$$0.6 > 0.5 > 0.4$$

Therefore, the EUD in this scenario is 1 (in [arbitrary] tenth of an expected point units). The other version of the PLT we are interested in has the following probabilities of reward (the mapping of keys to doors is identical): P(reward UL)=0.8, p(reward UR)=0.9, p(reward BL)=0.1, p(reward BR)=0.2. This results in an EUD of 3. Brown and Lovett (in preparation) have found that participants find it very hard to learn how to solve the game when the EUD is 1, while many more participants are able to learn the solution when the EUD is 3. The onset of learning is also much earlier in that condition. Interestingly, this effect is independent of the specific probabilities that are used to form the EUDs (Brown & Lovett, 2001). This justifies our decision to regard the EUD 1 and 3 versions of the task as two truly distinct conditions, and to compare the behavior of the model to human behavior under exactly these two conditions.

Description of the Model

Like participants, our model for learning probabilities and making choices in the PLT starts out with very little knowledge. The only initial constraint on its selection is the required mapping between keys and doors of the same color. Like participants, it will be making 200 choices, and learn about the probabilistic reward structure of the system during the process of making the choices and from the feedback following them.

At each "choice point", there are two available options (two doors matching the color of the key). The model selects the option that has a higher current *evaluation*. The following three factors contribute to the overall evaluation of each option:

(1)   The current estimate of the probability of receiving a reward upon opening that door.

(2)   The current estimate of the value of the key that will be given by that door, weighed by a parameter relating

the importance of future rewards to that of immediate rewards.

(3) A measure of *information gain*, that expresses how much knowledge of the characteristics of all available options, i.e. of the system as a whole, will increase as a consequence of "opening that door". This is weighed by a parameter that grades the importance of *information gain*.

A formal definition of the model's selection S at time t is:

$$S_t = \max(E_{it}, E_{jt}),$$

where E_{it} and E_{jt} are the evaluations of the two available choices i and j at time t. The evaluation of both options is analogously computed as

$$E_{it} = \frac{successes_{it}}{successes_{it} + failures_{it}} + k * \max(A_t; \frac{B_t + C_t}{2}) + c * N_{it}^{-.5}$$

The first term is the estimate of the probability of success of door i at time t. The second term is the value of the key given by the door, weighed by the parameter k. The value of a key is the estimate of the best future rewards that can be expected from having that key. This is either the current estimate of the repeatable cell whose door matches the current key (denoted by A), or the expected value of alternative sequences that starting with the current key but will give a different key (denoted here by the average of the expected value of the door that can be opened with the current key and will itself yield a different key (B) and the current value of that different key (C). The third term, finally, denotes the *information gain* associated with choosing door i at time t, weighed by the parameter *c*. This measure simply decreases as a power function of N_{it}, the number of times door i has already been chosen at point t. This expresses the assumption that we learn something about the system each time we make a choice (which is true, because we get feedback), but that this information shows marginally decreasing returns. The selection of a power function to represent this effect quantitatively is partly based on the fact that the posterior distributions of the reward probabilities associated with the four doors are beta distributions, the variance of which decreases as a power function of additional observations. This characteristic of the posterior probability distribution of an event again points to the crucial interaction between exploration and exploitation we are trying to capture here: the accuracy of the estimates, both of the reward-probabilities and the value of the keys, critically depends on sufficient exploration, more specifically: a sufficiently large number of observations. However, the impact of exploration does not remain constant, but instead decreases systematically: rapidly in the beginning, marginally later on.

The model presented is a learning model in the sense that all estimates are updated after each choice, as is the information gain measure associated with that choice. It does not use noise; the only sources of variability are changes in the estimates, which in turn are caused by the probabilistic feedback, and changes in the *information gain*

measure, which leads the model to abandon familiar options and explore less familiar ones.

The parameters in the model (*k, c*) are basically free parameters that can be adjusted to reflect a different impact of future rewards (*k*) or Information Gain (*c*), either under different circumstances or even between (simulated) participants. However, in all simulations presented here, they have been kept constant throughout, with *k*=3 and *c*=4.

Some Basic Predictions and Mechanisms

As the relative impact of the two competing components of the evaluations changes with time, the learning and behavior of the model can be describes as follows (note that they are not assuming discrete stages but continuous changes, the segmentation in the following paragraph was made to serve clarification):

(1) *Pure exploration*. Choices will be made based on the Information Gain measure, i.e. choices that have not yet been explored will be explored. When *c*=4, the estimates of the actual payoffs (which are initialized to 0.5) are still too small to counteract this during the first few cycles (i.e. until a sufficient number of experiences has been accumulated).

(2) *Early, inaccurate Estimates*. Because the measure of Information Gain decreases relatively rapidly in the beginning, the actual estimates of the reward probabilities and the key values begin to impact the models choices. However, due to the probabilistic feedback, and the fact that they are still based on relatively few observations, the estimates might not reflect the true ranking of the options, particularly in conditions where EUD=1. Two forces counteract these inaccurate representations. First, by choosing the currently best option repeatedly, the model gathers information that corrects its estimate. Now previously lower-ranked options can compete again for selection. Second, by exercising its bias towards the less explored options, it obtains a more accurate estimate of their payoff probabilities as well. Consequently, the model quickly recovers from initial false estimates. Now, the model has established estimates of the reward probabilities that are robust enough to remain unfazed by the occasional "failure"-feedback. The model will then consistently chose the option with the highest estimated long-term payoff

(3) *Familiarity breeds contempt*. Flexibility is maintained for a while beyond this point, because, as one option, or a combination of two options, is chosen repeatedly, its *information gain* measure decreases, while that of the other options remains constant. Thus, there is the chance that the model abandons an option again in favor of another, particularly if this competitor has a similar expected payoff. It is clear how appropriate this behavior is in the task described here, where the EUD between choices are sometimes as close as 1 unit, and the chance to "accidentally" settle on the wrong solution is pronounced.

(4) *Optimal choices with Intermissions*. Eventually, the model will learn to solve the game, because its estimates are becoming more and more accurate. However, because of the dynamics described in the previous section, the model will continue to explore alternative options. The Intervals between these "exploration fits" also follow a Power Function: the number of experiences between "exploration

fits" becomes much longer each time, until they are so widely spaced as to be without any relevance anymore.

We will examine how the model learns and behaves under different EUD-conditions of the task described here in the next section.

Comparison Between Model and Data

The data to which we compare the behavior of our model were collected by Brown and Lovett (in preparation). A total of 80 participants worked on the version of the PLT described in this paper (for related work using a deterministic version, see Brown & Lovett, 2001). 60 Participants worked under the EUD 1 condition (participants were collapsed from different groups that used different sets of probabilities to form an EUD of 1, because the specific probabilities had no effect on behavior, as reported by Brown & Lovett, 2001). 20 Participants worked under a EUD = 3 condition. The reward probabilities for both groups are given in Table 1.

Table 1: Reward Probabilities in the EUD1 and EUD3 conditions.

	UL	UR	BL	BR
EUD=1	p=0.6/0.7/0.8	p=0.8/0.9/0.9	p=0.2/0.3/0.5	p=0.4/0.5/0.6
EUD=3	p=0.8	p=0.9	p=0.1	p=0.2

Participants' choices in both groups were recorded on a trial by trial basis. Choices of UL and BL, which are the choices that yield the "better key" (red) were coded as "1", to indicate subjects' attention to the keys, as opposed to the immediate rewards (which would favor UR and BR, respectively, choosing these options was recorded as "0"). Based on the binary raw data, the proportion of "choosing left" was computed for each of the 200 trials. Note that only a repeated choice of UL can lead to a proportion > 50%. An increase above that level is thus indicative of choosing the solution, UL, more often than not. Finally, the proportions were averaged over ten-trial blocks, resulting in 20 ten-trial blocks that depict the aggregated learning trajectory for all participants. Additionally, response latencies were obtained for all trials.

It was our goal to have the model produce data of the same format as the empirical data. To this effect, we implemented it in a spreadsheet and recorded each of the 200 choices it made per run, coding them the same way the human data were coded. We produced 3*20 model runs under the EUD1 condition, using the probabilities given in table 4, and 60 model runs using the EUD3 condition, also using the same probabilities. The results of these model runs were aggregated in the same manner as the human data. The model (obviously) was constrained to open doors with the appropriate key in the same manner as humans were. Its choices were based on the evaluations elaborated in the preceding section.

We will first present and comment on the correspondence between model and data for both EUD conditions on the aggregate level. However, the quality of a model can also be assessed by determining how well its individual runs resemble the individual learning trajectories of participants, particularly when behavior is very variable, as is the case here. Therefore, we present some comparisons between individual participants and individual model runs that show that the model can produce a range of behavior consistent with that shown across the sample of participants. A note on the parameters: The values of k=3 and c=4 were chosen to obtain a good the fit to the data in EUD1. They were kept constant for all other comparisons, including the ones on the individual level.

EUD1 and EUD 3: aggregate learning trajectories

It is easy to see how the behavior of the model differs under different EUD conditions. When the EUD is 1, i.e. very low, the model needs more trials to arrive at estimates that are accurate enough to warrant exploitation of one option. However, the small advantage associated with the winner, combined with the model's bias towards exploration, limits the stability of behavior under this condition. While the model will eventually converge to solving the game even under this condition, it is, like humans, often unable to do so within the 200 trials allotted in the experimental task described here. However, if the EUD is as high as 3, learning is considerably sped up: the accuracy of the estimates increases faster, because the variability in binary feedback is lower as the probability of success (or 1 vs. 0) becomes more extreme. and the gap between the two options' estimates increases faster. Exploration continues to be beneficial under this condition, but is more rare, as the differences between the estimates are pronounced enough to lead to appropriate exploitation and to counteract the impact of the *information gain* measure.

Figure 2 shows the (aggregated) learning trajectories of data and model under conditions 1 and 3. The difference between the groups is captured nicely by the model, especially the logarithmically shaped learning curve in the EUD 3 condition. Note that neither humans nor model arrive at 100% exploitation of UL under EUD 3. This reflects (in the model) the response to the probabilistic feedback and the ensuing recurrent, brief "exploration bursts". Note particularly the downward dip in the final four 10-trial-blocks that is shared by empirical and model data. In the model, this is the consequence of having chosen the solution, UL, a considerable number of times. As a consequence, its *information gain* Measure has decreased sufficiently to allow the competitors a few more explorations: familiarity breeds contempt. Neither humans nor model show much learning under the EUD 1 condition, for the reasons outlined above.

An even more powerful test of the validity of our model, in particular our claim that a component of *information gain* is essential for understanding and predicting human behavior, is a comparison between the data and a version of our model that sets c=0, thereby completely eliminating the component of *information gain*. The results of this comparison are shown in Fig. 3.
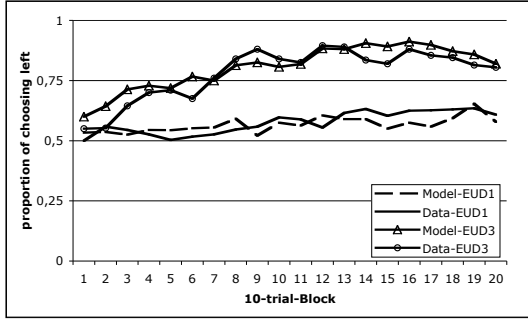
Figure 2: Model and Data Curves under EUD1 and EUD3 conditions.

The "no-info-gain" model does not capture the learning under the EUD3 condition. Levels of "choosing left" remain constant during all trials for this model. Even more importantly, the "no-info-gain" model operates without noise, and therefore does not exhibit any variability over time. The fact that the proportion of choosing left remains below 100% for the no-info-gain-model is an artifact of this: Some model runs *always* choose UL from the beginning, others *always* choose ER-BL throughout all trials. No changes occur, because none of these options has a bad enough payoff to jolt the model out of its inertia. The same inflexible behavior is true for the EUD1 version of the no-info-gain-model.
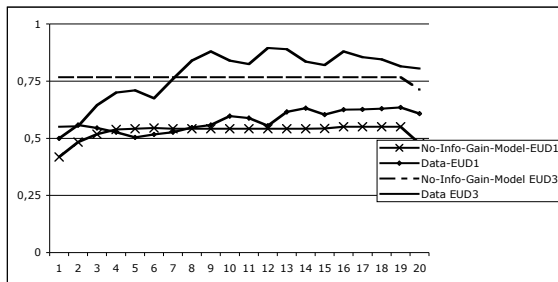

Figure 3: Predictions of a model without *information gain*.

This begs the question whether the addition of noise to the evaluation mechanism wouldn't have the same effect as the notion of explicit information gain. This is still an issue for future exploration, especially since there are two kinds of noise that can play a role here: The estimates themselves can be noisy, or the selection process that operates on them can involve noise, the subtle difference between these two concepts of noise, and possible integrations with the present model are the objective of future work. Here, the following argument can be made against the use of noise and in favor of the notion of *information gain* proposed here. Cognitive models of probability matching within the ACT-R framework have to assume an extremely high level of noise in order to capture the observation that there is still variability in participants behavior after a large number of experiences. We will show examples of this kind of behavior in the current task in the next section, and show

how our model can reproduce this without assuming any noise. We will address this issue again in the discussion.

Individual Participants and Individual Model Runs

Another measure of a model's quality that goes beyond the comparison of average curves involves the inspection of individual model runs with individual subjects. Especially in tasks that cause a high variability in behavior, this is interesting, because it enables us to inspect the flexibility of both model and humans. Therefore, we inspected whether we could identify individual model runs, in the set that fed the average curves shown in Fig. 2, that match the learning trajectories of individual subjects.

One characteristic of the model, both in the EUD 1 and the EUD3 condition, is that it can exhibit relatively "gradual" learning. Essentially, the model settles on a "current best" based on the current estimates, and is drawn away from it again by increasingly correct estimates (which perhaps reveal that the option it has settled on is not that superior after all, as well as the decrease in that option's *information gain* relative to the that of the other options. It gradually converges towards an increased choice of the true best, in this case UL. Figure 4 shows an example of this, the model being matched to participant 126.
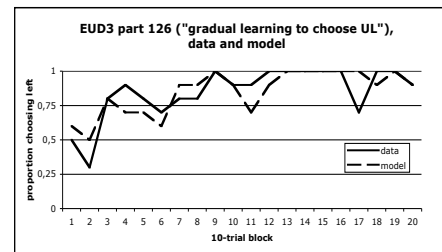

Figure 4

We can see similar patterns of learning in individual model runs under the EUD1 condition, even though the model does, on average, hardly learn. One striking and frequent pattern under the EUD1-condition, and one that we believe is hard to capture by a model that uses only noise as variability-inducer, is the complete abandonment and later re-uptake of the option UL. This pattern, which we call spikes (one of many examples is shown in Fig. 5) is due to two factors: Firstly, the quality of UL as "best" choice is less clear in EUD1 that in EUD2, so its estimate will remain closer to that of its competitors, occasionally falling below them. Secondly, the advantage that UL might have over the other options in terms of expected payoff is not big enough to counter the fact that its *information gain* measure will decrease as it is chosen more often, falling below that of the competitors: Familiarity breeds contempt. These two factors taken together model the fact that, under EUD1, the model can't establish sufficient "trust" in an option in order to exploit it: As it repeatedly chooses UL, its estimate remains mediocre, at the same time, the other options begin to seem more attractive again. If there is one option that has been explored particularly rarely, this option will promise a high *information gain and* will be chosen for a couple of trials,

until it has been established that its current estimate is below the UL and its *information gain* has decreased. The result is a "spike", as the one seen below. Fig. 6 shows one of the many examples of this pattern that can be found in the data and in the set of model runs. Here, one model run is matched to participant 293.
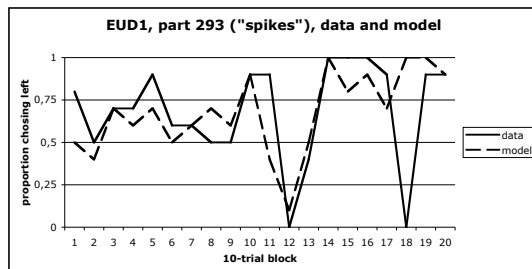

Figure 6.

Discussion

It has been our goal to demonstrate that Information gain, not just actual payoffs, can drive Decision Making in a probabilistic environment. To this end we have created a model that learns to make choices in a probability learning task, choosing among options based on its estimates of their actual payoffs and the *information gain* associated with selecting that option. Upon detecting that this model fits human behavior much better than a model that ignores *information gain,* we ask ourselves two questions: Is the behavior of such a model rational? And: Is the model correct? The answer to the first question is, in our opinion, a clear "yes". Exploration is appropriate in probabilistic environments, because it increases the accuracy of the probabilistic representations. The model, like humans, is able to adjust its amount of exploration to the structure of the probabilities, abandoning exploration early when they are easier to discriminate. Its need for exploration prevents the model from being stuck with the wrong choices, but this need is also systematically related to the structure of the environment, such that its impact will be smaller the fewer data are needed to arrive at reliable estimates.

But: Is the model as it stands now correct? This is unlikely. It has only recently been formulated and has only been tested with the datasets reported here. We regard the present version of the model as a skeleton containing the elements we believe are essential for explaining human choices and learning in a probabilistic situation. However, as the word "skeleton" suggests. augmentation is clearly called for. For instance, it is likely that the value of exploration, i.e. of information gain, varies from situation to situation. It is even more likely that Humans themselves can adapt its importance to different situational demands. Essentially, this calls for modeling work regarding systematic changes in the c parameter. Another open question concerns the precise definition of Information Gain. Right now it is a "raw", content independent indicator of how much we have learned by making a choice, and it always decreases according to the same function. This is a strong assumption, which must

be tested empirically. It remains to be seen whether the monotonous decrease of Information Gain remains adequate to model behavior in situations with, e.g., non-stationary probabilities. Human adjustment to this additional complexity will certainly pose another challenge.

Finally, the view put forth in this paper is that choices in probabilistic environments can be influenced by the explicit, active wish to explore. This notion is partially at odds with models that only assume a noisy estimation, or a noisy choice process in order to account for variability in behavior. A comparison between these two approaches can be resolved on two different levels: Experiments can be designed in which the two models make clearly distinct predictions, and formal analyses of both model can be conducted to reveal how the two models might "fit" different situations, and under which circumstances their choices converge. These efforts might lead toward a more complete theory of how these two drives of exploration and exploitation might be interacting in driving human behavior, an endeavor of which the model reported here merely scratches the surface.

References

Anderson, J.R. & Lebiere, C. (1998). *The Atomic Components of Thought.* Mahwah, New Jersey: Erlbaum.

Broeder, A. (2000). Assessing the Empirical Validity of the "Take The Best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26(5),* 1332-1346.

Brown, J.C. & Lovett, M.C. (2001). The effect of reducing information in a modified Prisoner's Dilemma Game. In J.D. Moore & K. Stenning (Eds.), Proceedings of the 23rd Annual Meeting of the Cognitive Science Society (pp. 162-167). Mahwah, New Jersey: Erlbaum.

Brown, J.C. & Lovett, M. C. (in preparation). Learning to Choose in a Non-Deterministic, Single-Player Version of the Prisoner's Dilemma Game. Carnegie Mellon University, Pittsburgh, PA.

Estes, W.K. (1964). Probability Learning. In A. W. Melton (Ed.), *Categories of human learning.* New York: Academic Press.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is relevant for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective Probability.* New York: Wiley.

Payne, J.W., Bettman, J.R. & Johnson, E.J. (1993). *The Adaptive Decision Maker.* New York: Cambridge University Press.

Rieskamp, J. & Hoffrage, U. (2000). *When do people use simple heuristics and how can we tell.* In G. Gigerenzer, P.M. Todd & the ABC Research Group, *Simple Heuristics that Make Us Smart.* New York: Oxford University Press.