

**LEARNING IN THE ABSENCE OF FEEDBACK –
AN EXPERIMENTAL STUDY**

Christina Fang
Department of Management
The Wharton School
2000 SH-DH
3620 Locust Walk
Philadelphia PA 19104
cafang@wharton.upenn.edu

April 28, 2003

DRAFT: PLEASE DO NOT CITE.

I am grateful to Paul Kleindorfer and Howard Kunreuther for generously funding this research through the Risk Management and Decision Processes Center at the Wharton School. I would like to thank Sachin Rekhi for many enlightening discussions and excellent programming help. Daniel Levinthal, my thesis advisor, provided much guidance in this work and throughout my dissertation. I am also grateful to David Croson, Rachel Croson, Jerker Denrell, Anne Marie Knott, Jen Yue Shang, Nicolaj Siggelkow and Sidney Winter for helpful comments and suggestions.

Abstract

Sequences of decisions where the consequences obtained at one stage may determine the choice alternative available at the next one are known as dynamic decisions. Since a long series of decisions need to be made, one important feature of this type of decision is the lack of immediate outcome feedback. How can decision makers learn in the absence of outcome feedback?

In this paper, I report an experimental study of learning in the absence of immediate feedback. The setting is a computerized treasure hunt game, where the treasure can only be found at a fixed location. Results indicate that subjects seem to develop a mental map of the relative values of different locations, starting from the locations that are closest to the solution.

1. Introduction

Sequences of decisions where the consequences obtained at one stage may determine the choice alternative available at the next one are known as dynamic decisions (Einhorn and Hogarth, 1981). Since a long series of decisions need to be made, one important feature of this type of decision is the lack of immediate outcome feedback. The challenge is therefore – how can decision makers learn in the absence of outcome feedback?

If the decision maker was a student of rational choice perspectives, the general form of the problem and its normative answer should be quite familiar (Meyer and Shi, 1995). It is a well-known variation of the armed bandit challenge, a name which stems from the toy problem of a gambler trying to choose between two slot machines with uncertain odds (DeGroot 1970). As a result, dynamic programming techniques (Bellman, 1957), real-options analysis (Dixit and Pyndick, 1994) and optimal statistical theory (DeGroot, 1970) become the natural and appropriate solutions. However, these approaches depend upon the ability to specify the full decision-tree of possibilities. In particular, they suffer from what Bellman (1957) termed the curse of dimensionality as space of possible futures expands exponentially with possible contingencies.

Nevertheless, these optimal theories may well provide the tools to describe the task environment of a sequence of inter-related decisions. While technically this problem is hard to solve, the dynamic programming formulation is both intuitive and insightful. More importantly, its formal calculus is promising for a psychological interpretation (Brehmer 1990, 1995; Gibson, Fichman and Plaut, 1997).

In my dissertation (Fang, 2003), I examine one plausible model of learning known as the credit assignment model (Sutton and Barto, 1988). It incorporates the basic principles of dynamic programming into a computer algorithm. Using computer simulations, I show

that firms can develop alternative basis for learning by constructing their own mental models (Thagard, 1996). These mental models emerge as firms learn to associate rewards with past actions according to the temporal/structural proximity between the specific past action and the eventual outcome, much akin to the law of the effect (Thorndike, 1911). In other words, firms are modeled as gradually assigning the "credit" arising from the overall sequence of actions to each of the antecedent actions. First, states that are closer to the solutions are recognized as valuable as credit is assigned to them initially. As credit cascades to states that are further away, firms develop a more fine-grained cognitive structure. I show that learning this way improves performance as a problem solving task is repeated.

In addition, I also highlight several potential traps of learning, in particular when excessive attention is paid to proximate subjects in the decision sequence. By manipulating the depth of credit assignment, i.e. the number of states in the temporal sequence to which credit is assigned, I show that firms with a deeper depth learn faster initially, though their long term performance tends to be inferior. The reason is that firms with a deeper credit assignment sequence tend to learn excessively from the initial noise inherent in their first hits. This is consistent with 'superstitious learning', as decision makers learn a mis-specified correlation between outcome and action.

In this paper, I report an experimental study of learning in the absence of immediate feedback. It is motivated by a simple question – is the credit assignment model a realistic description of human behavior? In other words, do human beings, confronted with a task without immediate outcome feedback, resolve it in the prescribed manner? In particular, I test the two predictions from computer simulations discussed above.

2. Method

2.1 Experimental design

In order to study how human beings make decisions in the absence of immediate outcome feedback, I design a computer-based treasure hunt game. In this simplified treasure hunt game, the space that contains the treasure is a collection of rooms that resembles a miniature globe. Subjects are required to hunt for the treasure which is only found in one of the rooms, whose location remains fixed throughout the experiment. Each room is uniquely identified by a sign with two defining characteristics – a shape and a color. In this context, outcome feedback is only available when the treasure is found; at every move in the space, the subjects are not informed about whether they are 'closer' or 'further' away from the treasure. The size of the globe is 6 by 6. As such, there are all together 36 rooms. The size of the space, however, is concealed from the subjects.

The actual hunt proceeds as follows. Subjects are first randomly started in the space and are assigned one of the rooms as their starting position. Subjects are allowed only local vision – they can view only those rooms that are adjacent to the room where they are currently in. In other words, all they can see from the computer screen is five rooms – the room in the center is their current locations, while four others rooms immediately adjacent to it are also shown, with their marking signs. As illustrated in the following diagram, subjects can only see the rooms north, south, east and west of their own.

Insert Figure 1a here

At each move, they need to decide whether to go north, south, east or west. Once that decision is made, the computer screen is changed to reflect their new location and vision. The chosen room now becomes the center room and a new set of corresponding

neighboring rooms is seen from the screen. For instance, if the subject chooses to go right, the right room with a circle will now appear at the center location as seen below:

Insert Figure 1b here

In this way, subjects move in the space until they find the treasure. Once the treasure is found, the screen will freeze – only the room where the treasure is and the room where subjects come from are shown. The three other adjacent rooms to the treasure are not shown. Five seconds will elapse before the second trial begins. In the next trial, the location of the treasure and the physical characteristics of the space remain fixed. However, the subjects will now be randomly started in a different room in the space and proceed as before.

Unknown to the subjects, they are randomly assigned to one of the two experimental conditions. Under the first condition 'Baseline', during the five-second interval between trials, two rooms are shown: 1) the room where the treasure is and 2) the room where the subjects come from. No other objects on the screen are shown. Under the second condition we called '2-step', in addition to the above information, the corresponding symbols of these two rooms are also displayed in the right hand corner of the screen, with their temporal sequence highlighted. This is illustrated in Figure 1c below. This manipulation aims to direct attention of the subjects to the existence of “1-step neighbor' and a ‘2-step neighbor’, and therefore manipulates the depth of the credit assignment sequence.

Insert Figure 1c here

In order to capture the subjects' valuation of various states i.e. rooms which are more or less proximate to the treasure, I collect this information at fixed intervals. After every five

trials, the game is stopped and an online message box pops up on the screen. The subjects are asked to rank order four different rooms in terms of their perceived proximity to the treasure and therefore value. Unknown to the subjects, the four rooms are chosen such that they are 1-step, 2-step, 3-step and 4-step away from the treasure respectively. This way, we obtain a qualitative sense of subjects' relative valuations. Once the subjects finish this evaluation, a new trial begins.

At the end of all the 20 trials, I implement an online post-experimental Q&A in order to gauge the subject's explicit knowledge about the game itself. I ask subjects a couple of open-ended questions like the kind of strategies, if any, that they have been following and what advice they might give to others who are about to play this game. Subjects can choose to write in whatever amount of detail s/he thinks fit.

2.2 Set up

A total of 68 subjects are recruited, who participate in the experiment in 9 sessions about 7 subjects per session. Upon arrival at the laboratory, the subjects are seated in a single room and are handed written instructions and informed consent forms. The experiment is computer-based. After the instruction is read to the subjects, the experiment is run on a PC. It consists of 20 trials of a treasure hunt exercise for each subject. Subjects will be randomly assigned to one of the two experimental conditions – the 'Baseline' and the "2-step". About 36 subjects happen to fall into 'Baseline' and another 32 subjects in '2-step'.

2.3 Performance measures

In addition to partial course credit, subjects are also paid according to their performance in the experiment. Performance is measured by two indicators – one denotes the total number of moves (rooms) the subject has visited and the other denotes the total amount of time s/he has used up to find the treasure since the beginning of each trial.

Subjects are told to minimize **both** the number of moves **and** the total time. In addition, the **average** of their performances across all 20 trials is then compared to that of the other subjects. Each subject will receive the following compensation depending on which bracket his or her performance fall in comparison with the overall performance:

Top 10%	\$10
Next 35%	\$7
Next 35%	\$5
Bottom 20%	\$0

The choice of two performance indicators is first motivated by the fact that subjects, unlike computer agents, necessarily take different amount of time to carry out one move. In addition, if performance is only measured by one of the two indicators, e.g. total time to solution, then one viable strategy to optimize performance is to move as quickly as possible in the hope of randomly hitting upon the treasure. If ‘navigating’ and ‘moving in the space’ is assumed to be costless, this constitutes the optimal strategy. On the other hand, if performance is solely measured in terms of number of moves, subjects may be overly cautious and end up spending too much time pondering over the movements. Neither behavior is relevant for the purpose of our study. As such, the experiment is designed to purposely discourage and possibly eliminate such ‘strategic’ behaviors outside our realm of interest. To do so, subjects are instructed to minimize **both** indicators. Furthermore, we choose to be deliberately vague about the weights attached to each indicator so as not to direct attention to some potential tradeoff between the two indicators.

In addition, subjects are told that their final performance is determined by taking average performance across all 20 trials. Since early trial performances do matter, I encourage the subjects to be mindful about striking a desirable balance between

exploration and exploitation. An alternative incentive design is to measure their performance only at the very last trial. This design encourages subjects to explore extensively for the first 19 trials before switching to the mode of exploitation at the very end. While it may be interesting to investigate learning behavior under this alternative specification, I choose the existing design primarily because it allows us to examine learning behavior across trials.

3. Analysis and results

Figure 2a and 2b plot average performance in terms of both average total time to treasure and average number of moves to treasure respectively as a function of trials under both conditions. First, it is noteworthy that in both cases, there is rapid decline in both time and moves at the very beginning of the game. Performance improves drastically and quickly. Subsequently, performance becomes more stabilized and the rate of improvements slows down considerably. Consistent with the model predictions by computer simulations, human subjects indeed exhibit substantial improvement in performance over time.¹

Insert Figures 2a and 2b here

While this dramatic improvement of performance clearly provides prima facie evidence of learning, it does not however reveal the kind of mechanism through which learning takes place. The central tenet of this dissertation is that learning proceeds by a

¹ While this is clearly evidence of learning, the performance curve after log transformation does not correspond exactly to the conventional learning curve. After transformation, the curves do not resemble a straight line. Instead, they have a lot of curvature, similar to Figure 1 and 2. This implies that performance is improving even faster than the exponential rate.

process of credit assignment, whereby subjects develop mental models of the values of states more or less proximate to the solution (Fang, 2003). In order to fully test this model, I need to, at the very least, reveal an underlying mental model that increasingly better differentiates states more or less proximate. In addition, I also need to examine whether varying the depth of credit assignment sequence matters.

I begin by investigating the latter. Computer simulations have shown that more credit assignment as a result of a deeper and longer sequence results in faster learning in the very beginning by reducing the amount of random search. However, longer sequence also results in inferior longer term performance as it tends to repeat random paths tried earlier. Manipulating the depth of credit assignment sequence is achieved in the experimental treatment of ‘2-step’. Recall that in this treatment, subjects are shown a 2-step sequence of rooms they have visited just before hitting upon the treasure. The treatment therefore directs subjects’ attention to the two rooms that have led them to the treasure. As such, it manipulates the subjects’ conscious memory of various lengths of the path they just take and therefore the depth and amount of credit assignment they carry out.

In Figure 2a, there is a drastic decline in number of moves from trial 1 to trial 2. For these two early trials, it is noteworthy that the darker line (representing the performance of subjects in the 2-step treatment group) is well below that of the baseline case. As such, subjects in the treatment group are able to find the treasure in substantially less number of moves, at least in the initial phases of the experiment. In addition, this is not associated with more time to solution, as seen in Figure 2b. This suggests that at least for the very first couple of trials, a deeper credit assignment sequence (as in the treatment group) does seem to help improve performance at a faster rate than the baseline case.

This is in line with the simulation prediction that having a longer path in memory helps to reduce the amount of randomness in the beginning of search. This window of opportunity, however, is very short. For all subsequent trials beyond the second, while the number of moves to find the treasure remains similar, subjects in the treatment group generally take **more** time. As predicted by the simulation model, subjects who use a longer credit assignment sequence tend to have inferior performance in the longer term. Always mindful of a 2-step sequence, they may end up looking for a 2-step sequence that works in the past while a more efficient strategy may be simply to explore a little bit more. However, subjects in the control ‘Baseline’ group do not have the luxury of additional information. This ‘lack of information’ forces them to explore more and exploit less. As a result, despite an earlier disadvantage, subjects in the control group are able to outperform slightly those in the treatment group for the majority of the trials.

Since subjects are started randomly in each trial, performance as depicted in Figure 2 reflects differences in initial conditions. To further examine the extent of superstitious learning, I plot performance as a function of the starting distance for three phases of the experiment in Figure 3 and 4 respectively.

Insert Figures 3 and 4 here

As seen in Figures 3 and 4, there is generally a decrease in performance as the starting points of the trial increase from 1 to 6.² In Figure 3, which plots performance as indicated by the number of moves, as the game progresses, subjects in the group ‘2-step’ take more moves to find the treasure. This may be indicative of a more circuitous path as

² At distance of 6, there is much less number of observations. As such, the curvature of the performance curves around 6 is less reliably measured than that around other distance points.

subjects assign excessive amount of credit to earlier moves.³ In Figure 4, which plots performance as indicated by total time to treasure, the same trend is observed. Performance of subjects in '2-step' is generally inferior to that in the control group, though the differences seem to narrow toward the end of the game. By controlling for the starting points, I confirm that subjects who do more extensive credit assignment do well in the very beginning but not as well in the medium and long term.

In order to fully test the model of credit assignment, we need to reveal the underlying mental model of subjects. This is achieved by measuring and collecting their subjective valuation of the rooms at distances 1-4 away from the treasure. Figure 5a and 5b plot the average ranking of all four rooms across all subjects in the same treatment group over time as a function of the actual distance of those rooms from the treasure.

Insert Figure 5 here

Two interesting findings emerge. First, under both conditions, subjects learn accurately the relative valuation of 1-step vs. 2-step neighbors. Subjects seem to be able to identify earlier on in the game that a 2-step neighbor is in fact further away from the treasure than the 1-step neighbor. This is evident in the positive slope between 1 and 2 (along the x-axis) that is observed universally over time under both conditions. If this success extends to all four randomly chosen neighbors, the curves will have exhibited a positive slope throughout. However, this is not the case. Subjects seem generally unable to differentiate 3-step neighbors from those which lie 4-step away, which is not surprising given their limited experience. Nevertheless, the accurate learning of the relative ranking

³ Also, the performance curves exhibited by subjects in '2-step' tend to fluctuate more over distances while those by subjects in 'Baseline' tend to be more stable.

for states that are close by shows that at least for these states, subjects manage to develop a fine-grained sense of relative value.

Second, it is interesting to observe that over time, while the ranking curves resemble closely each other for subjects in the control group ‘baseline’, there is quite a lot of ‘reshuffling’ in the relative rankings by subjects in the ‘2-step’ treatment group. In particular, while up to trial 10, subjects in the latter group can only differentiate between 1-step and 2-step neighbors, they seem to become more informed from trial 15 onwards. The ranking curves for these subsequent trials show positive slopes not just between 1-step and 2-step; rather, the positive slope extends now to the 3-step neighbor. This shows that subjects can now intelligently differentiate among these states which are relatively close to the treasure. More importantly, this demonstrates that as subjects gradually learn about the value of more or less proximate states, this evolution of valuation proceeds in a “backward cascading” fashion. States that are closest to the treasure are discovered and valued first. Only then are states that are a bit further away recognized. As such, this provides at least partial confirmation and evidence that human beings may indeed learn in the manner prescribed by the credit assignment model.⁴

4. Discussion and future Work

This study seeks to investigate whether the credit assignment model is a realistic and adequate description of human decision making behavior in the absence of

⁴ It is however intriguing why such gradual learning does not take place for subjects in the control group who seem unable to improve their understanding of the relative ranking over time. Further thinking suggests that this is perhaps an artifact of the experimental design. In particular, the four rooms are randomly chosen without regard to whether the subjects have either ‘visited’ or ‘recognized’ them before. As such, it is highly likely that subjects find some of these rooms rather unfamiliar, which is partly confirmed in post-experimental debriefing and interview. To rank these rooms therefore requires at least some ability to remember and recall. Under the treatment condition, the task of literally remembering these rooms are indirectly made much easier since during each trial except the first one, the 2-step sequence is shown. As such, more rooms have been reinforced or made salient in subjects’ minds. This may perhaps explain why subjects seem able to improve their valuation accuracy over time.

immediate outcome feedback. Analyses of experimental result generally confirm that subjects are able to improve their performance over time. In addition, this learning is driven by increasingly fine-grained and intelligent mental models which recognize the value of states closest to the solution first. I further show that excessive credit assignment leads to over-exploitation of prior experience and inferior performance in the long run.

These arguments are echoed in subjects' answers in the post-experimental questionnaires. A number of findings emerge. First, when asked to describe their own strategies in the game, the vast majority of subjects report that they try to remember rooms that are closer to the treasure. As seen from the excerpts below (emphasis added), subjects frequently use words such as 'led to', 'close', 'proximate' and 'near' to denote their key strategy:

I looked for three signs that I recognized- the red square, the yellow square, or the blue blob...**all were one or two moves away from the treasure.** It was too hard to remember too many symbols, so I focused on a small number that I recognized. (Subject 105)

Eventually I realized that certain shapes **led me** to **certain parts of the globe** where the shapes that I familiarized with the treasure were located. Then when I saw those shapes I memorized a pattern and found the treasure rather quickly. (Subject 102)

I just remembered a few patterns that got me to the treasure and then I just remembered not to just go in a straight line. **Once I had hit on one of the patterns, I would then know my way there.** (Subject 903)

I tried to remember the shapes and colors located **near** the treasure. That way, toward the end of the game I did not have to think about where I had to go as much as I did in the beginning. (Subject 604)

My strategy was to go around the globe first and try to find the treasure room, once I found the treasure room, I remembered the signs for the **adjacent** rooms, after that I just looked for those signs in finding the treasure. (Subject 103)

I remembered the rooms **next to** the treasure and looked for them.(Subject 705)

To further understand the prevalence of this credit assignment, I also carry out a simple count. Whenever key words such as "close to", "lead to", "proximity", "near", "one-step away", "next to", appear in the answer, the strategy is coded 1 as a credit

assignment approach. These words are chosen as they generally indicate an awareness of the importance of neighboring states to the treasure. 61% of the subjects in the "Baseline" group adopt the credit assignment approach, as 82% of the subjects in the "2-step" group do.⁵ Subjects who do not follow a credit assignment approach seem to be guided by specific, randomly occurring combinations of colors and shapes.

Second, subjects also seem to be able to develop a mental map in a cascading manner, successively discriminate among states that are more or less proximate to the treasure. In the excerpts below, they demonstrate a clear awareness of the temporal and thus value sequence, starting from the rooms closest to the treasure.

I tried to remember the object that I clicked **before** the treasure was on the screen. Once I remembered that object **as an indicator**, I tried to remember the object that **got me to the indicator**. Eventually, I had a pattern. (Subject 508)

I kept looking for the symbols that were **directly related** to the treasure, for instance, in the first trial I found the green lightning bolt next to the treasure, so I kept looking for that. As the game progressed, I would look for **other** adjacent symbols. (Subject 503)

I remembered key room identities and their location relative to the treasure room. And **once** I got more comfortable with which rooms were **immediately adjacent** to the treasure room, I began to get familiar with which rooms were adjacent to those rooms. (Subject 805)

To double check whether subjects truly believe in their professed strategies, I also ask them predictive questions such as 'what kind of advice they will give to a friend who is about to do this experiment'. Subjects describe vividly the backward cascading nature of the mental map in excerpts below:

I would tell him to try to create a grand map of the globe in his head, beginning with which rooms are adjacent to the treasure room and slowly **spreading out** from there. (Subject 805)

Be careful to note the 4 symbols **surrounding** your treasure and specifically where your treasure lies **in relation to** each. I ended up using only 2 of those symbols effectively, and I would have done better if I were more cognizant of **other pathways** to the treasure. (Subject 203)

⁵ I code strategies conservatively. Only if subjects provide a generalized strategy description, would they be coded as 1. If they simply describe the rooms i.e. colors and shapes, that they have found useful, without indicating the common characteristics shared by these rooms, their strategies will be coded as 0s.

While these two findings are common to subjects in both the control and the treatment groups, "exploitation of past sequences" seems to feature more prominently in descriptions by subjects exposed to '2-step' manipulation. In some cases, subjects report that they choose to stick with past sequences knowing that these may not be the shortest paths. Random exploration is described as a 'mistake'.

I clicked on the figures that had previously taken me to the treasure and **stayed away** from those that I had not come across when finding it. (Subject 901)

I used the two closest pieces and **just** searched for them throughout the game. If I got different pieces by **luck** after that I used those also to help me (Subject 710)

After I found out what the treasure was, I used the two rooms **I knew** it was near, and remember those for the rest of the game. Then, in later rounds, if I would reach the treasure **by mistake**, I would remember those other rooms that were also close to the treasure. (Subject 807)

Finding shapes that I recognized from **previous trials** that had led me to the treasure before, **following similar paths** even though there may have been closer ways to get to the treasure, I **stuck with the similarities I could remember and were sure of.** (Subject 806)

Attempt to remember paths in your mind. **Although they may not always be the shortest route, once you remember a path you remember more symbols that directly lead to the treasure and you no longer spend time randomly searching for something you recognize.** (Subject 808)

These excerpts, taken as a whole, provide supporting evidence that human beings learn via successive attempts at credit assignment in the absence of immediate outcome feedback. There are however several caveats to this experiment. First, there is some ambiguity in the way performance is measured. While most subjects do not appear to be cognizant of any potential tradeoff between the two performance indicators, it is nevertheless a confounding factor that needs to be more carefully controlled. This may be achieved by limiting the number of moves that can be made within a certain time unit. Second, in order to uncover the underlying mental model of valuation, the study uses a rudimentary method of soliciting relative rankings. There is also no consequence of 'wrong' rankings since subjects' performance is not directly related to their valuation exercise. As such, the revealed mental model is at best a rough approximation of the true

underlying valuation. More sophisticated methods of preference revelation such as auctions may be employed to further this line of inquiry.

Despite these caveats, this study provides some indication that human beings learn via a process of credit assignment. In addition, it furnishes a platform or template on which future experimental investigation can be based. The existing set up of the experiment can be easily amended to examine several issues at the heart of strategy and technological innovation. For instance, it will be interesting to compare subjects' behavior if their performance is measured only at the very last trial. This allows us to investigate in a simple experimental setting, issue of inter-temporal adaptation and the balance of exploration and exploitation. A second interesting extension is to change the location of the treasure (with or without the explicit knowledge of subjects) at either fixed or variable intervals to study issues related to radical vs. incremental innovations. More ambitiously, future work may further delineate memory based explanation of learning vs. strategy-based ones (e.g. credit assignment).

References:

- Bellman, R. (1957). *Dynamic Programming*. New Jersey, Princeton University Press.
- Brehmer B. (1990). "Strategies in real time, dynamic decision making", in *Insights In Decision Making* edited by Robin Hogarth, University of Chicago Pres, Chicago.
- Brehmer, B. (1995). "Feedback delays in Complex Dynamic Decision Tasks". In P. Frensch and J. Funke (Eds.), *Complex Problem Solving, The European Perspective*. Hillsdale, NJ, Erlbaum.
- DeGroot (1970). *Optimal Statistical Decisions*. New York, McGraw-Hill Book Company.
- Dixit, A., & Pindyck, R. (1994). *Investment Under Uncertainty*, New Jersey: Princeton University Press.
- Einhorn, Hillel J. & Hogarth, Robin M. (1981). "Behavioral Decision Theory: Processes of Judgment and Choice." *Journal of Accounting Research*, Vol. 19 Issue 1, p1-32.
- Fang, Christina (2003). "Strategy as Valuation", unpublished doctoral dissertation, The Wharton School, University of Pennsylvania.
- Gibson, F., M. Fichman and D. C. Plaut (1997). "Learning in Dynamic Decision Tasks: Computational Model and Empirical Evidence," *Organizational Behavior and Human Decision Processes*, 71, p11-35.
- Meyer RJ and Shi Y. (1995). Sequential Choice under Ambiguity: Intuitive Solutions to the Armed Bandit Problem. *Management Science*, 41(5) p817-833.
- Sutton, R. and A. Barto (1998). *Reinforcement Learning: an Introduction*. Cambridge, MIT Press.
- Thagard, P. (1996). *Mind: Introduction to Cognitive Science*. Cambridge MA, MIT Press.
- Thorndike, E. L. (1911). *Animal Intelligence*. New York: Macmillan.

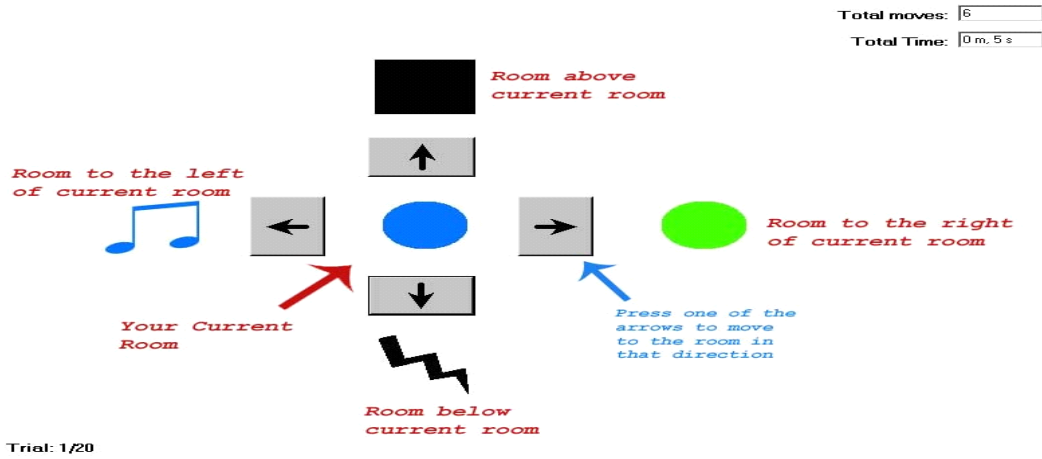


Figure 1a: Computer display before a move is made

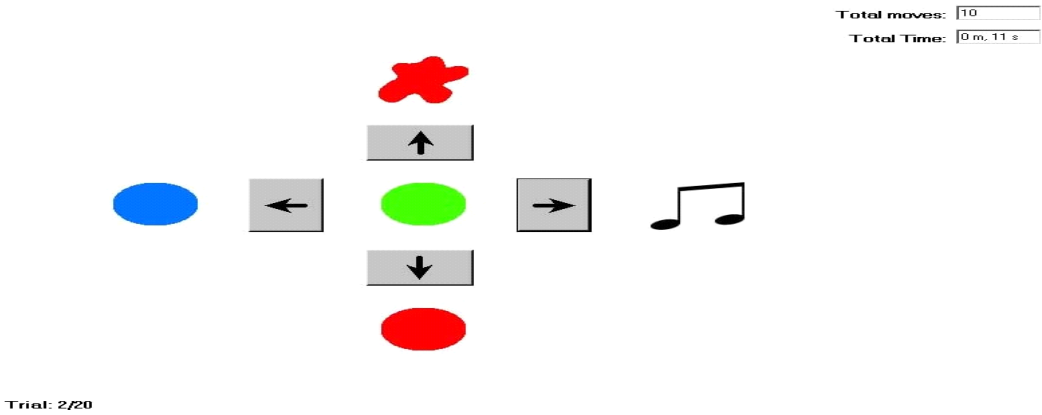


Figure 1b: Computer display after making a move

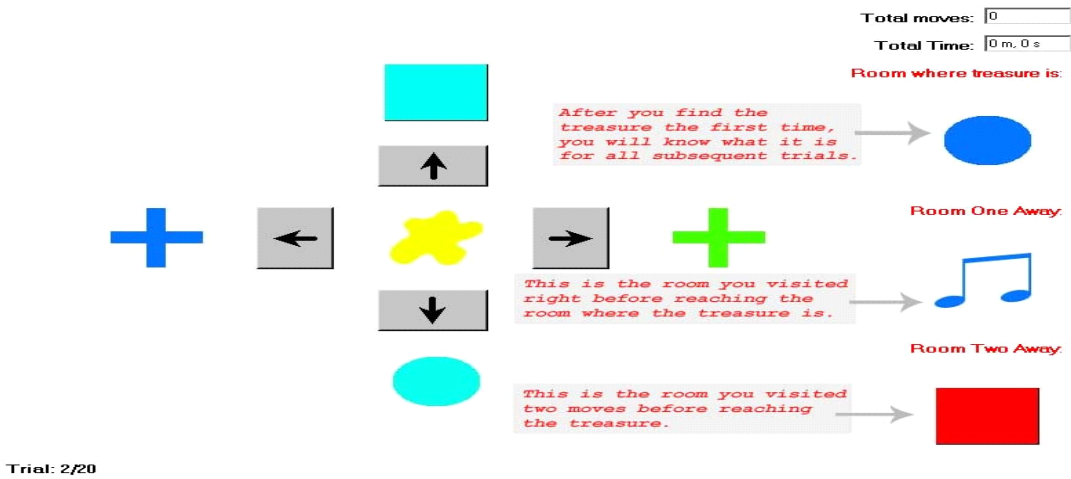


Figure 1c: Computer display for the treatment group '2-step'

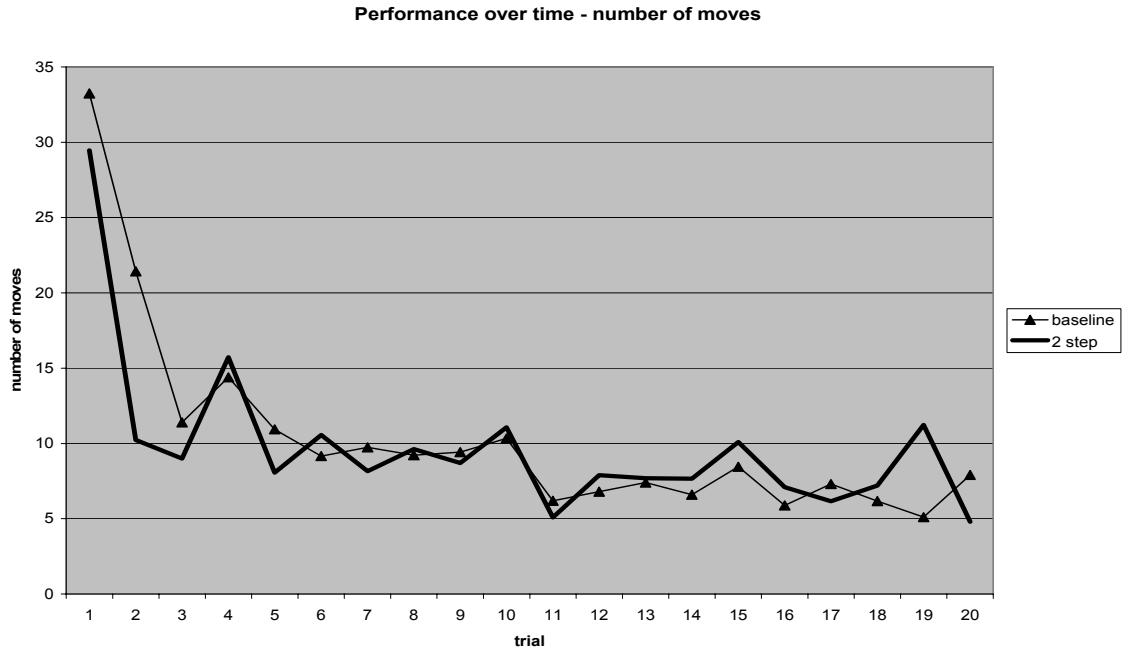


Figure 2a: Performance over time (# of moves) under both conditions

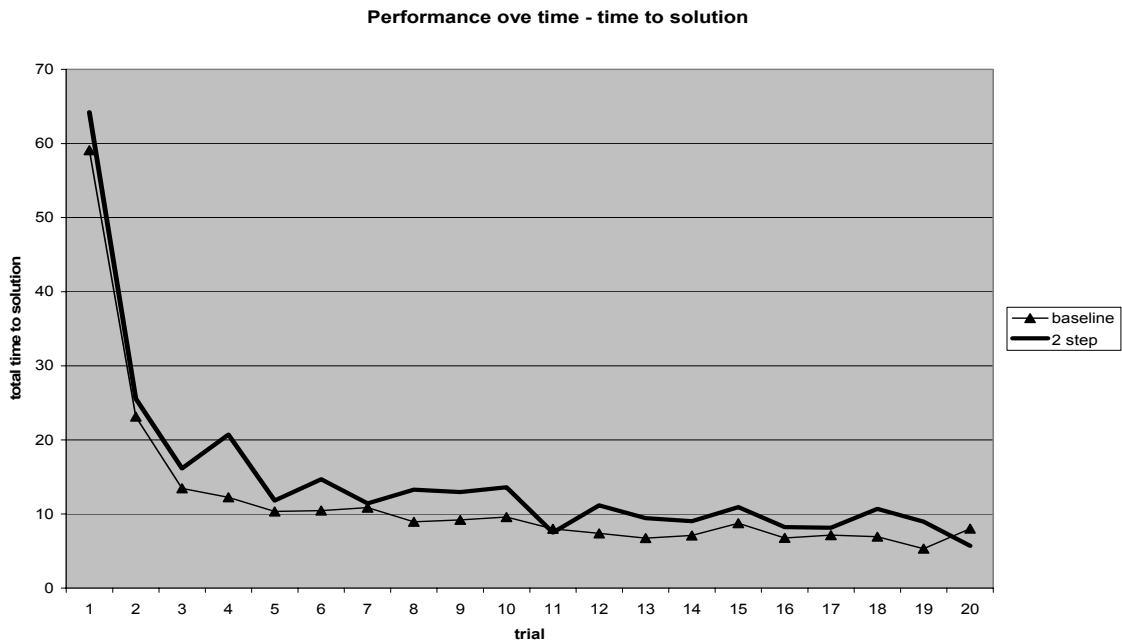


Figure 2b: Performance over time (total time to solution) under both conditions

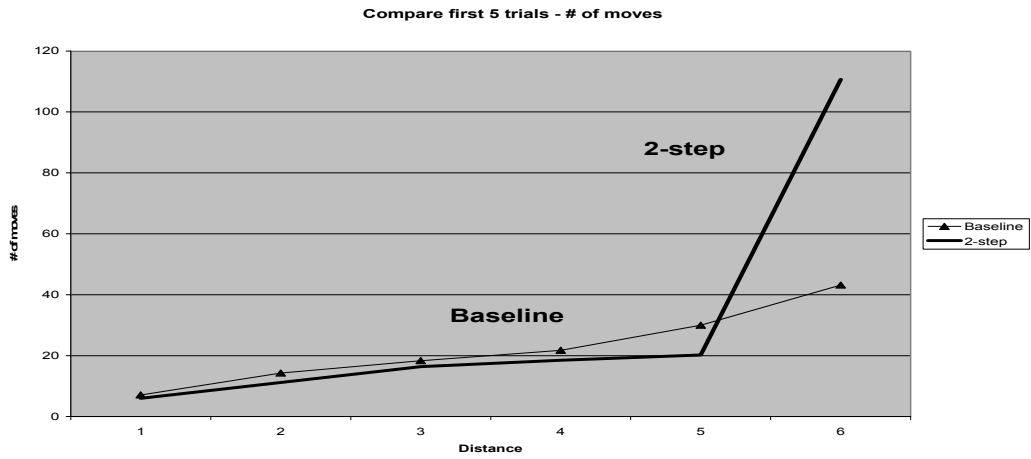


Figure 3a: # moves to treasure as a function of starting distance for trials 1-5



Figure 3b: # moves to treasure as a function of starting distance for trials 8-12

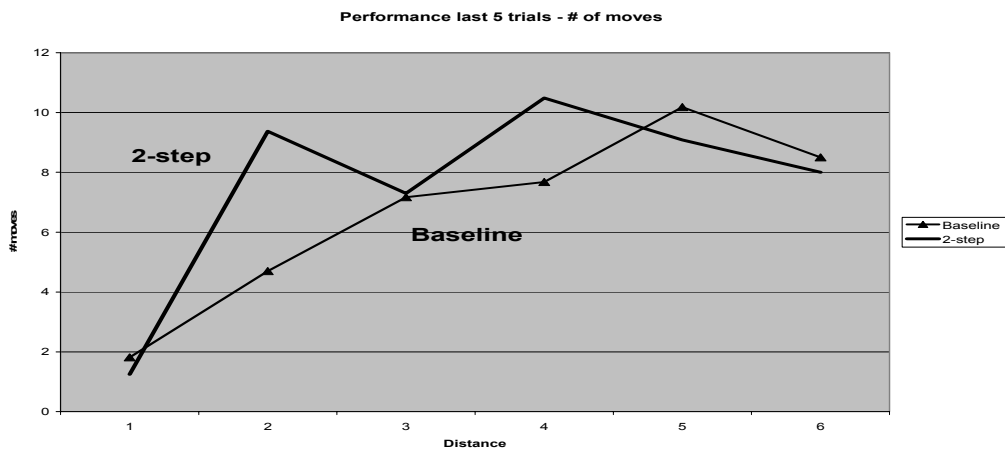


Figure 3c: # moves to treasure as a function of starting distance for trials 16-20

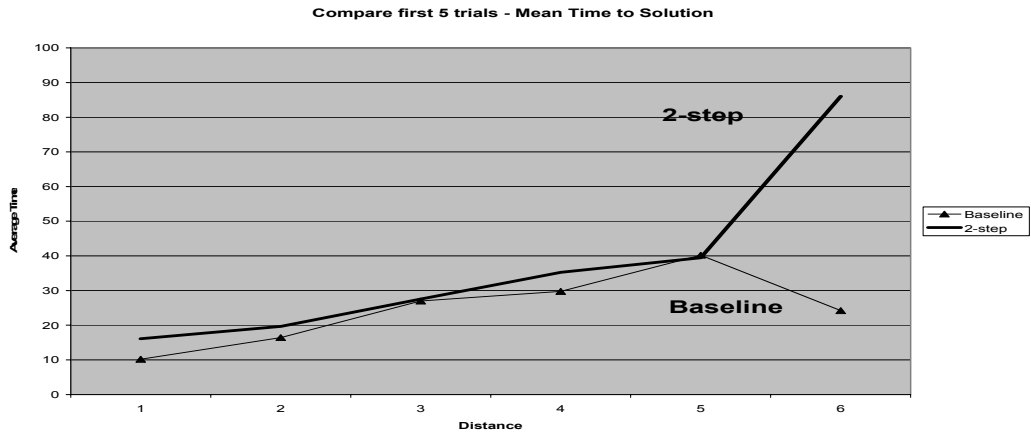


Figure 4a: Time to treasure as a function of starting distance for trials 1-5

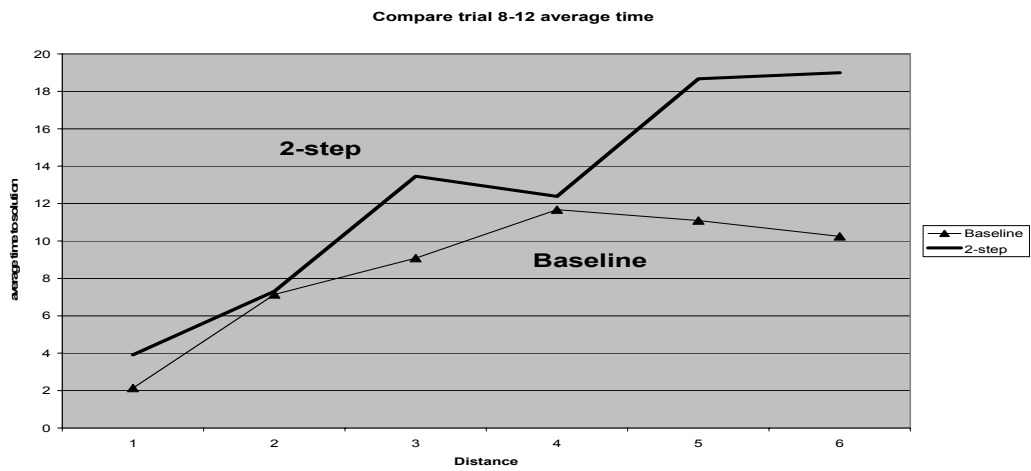


Figure 4b: Time to treasure as a function of starting distance for trials 8-12

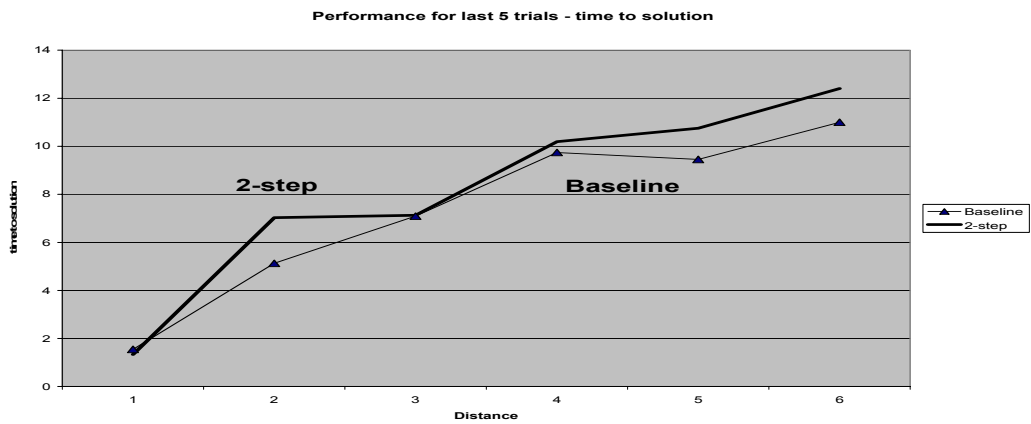


Figure 4c: Time to treasure as a function of starting distance for trials 16 to 20

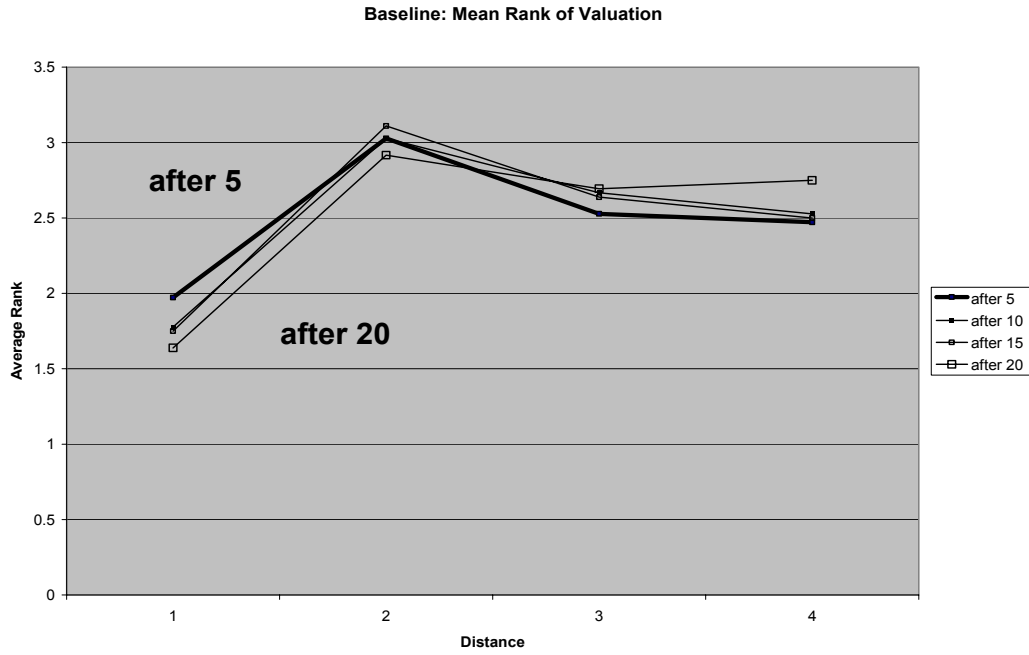


Figure 5a: Mean ranking of valuation for the control group 'Baseline'

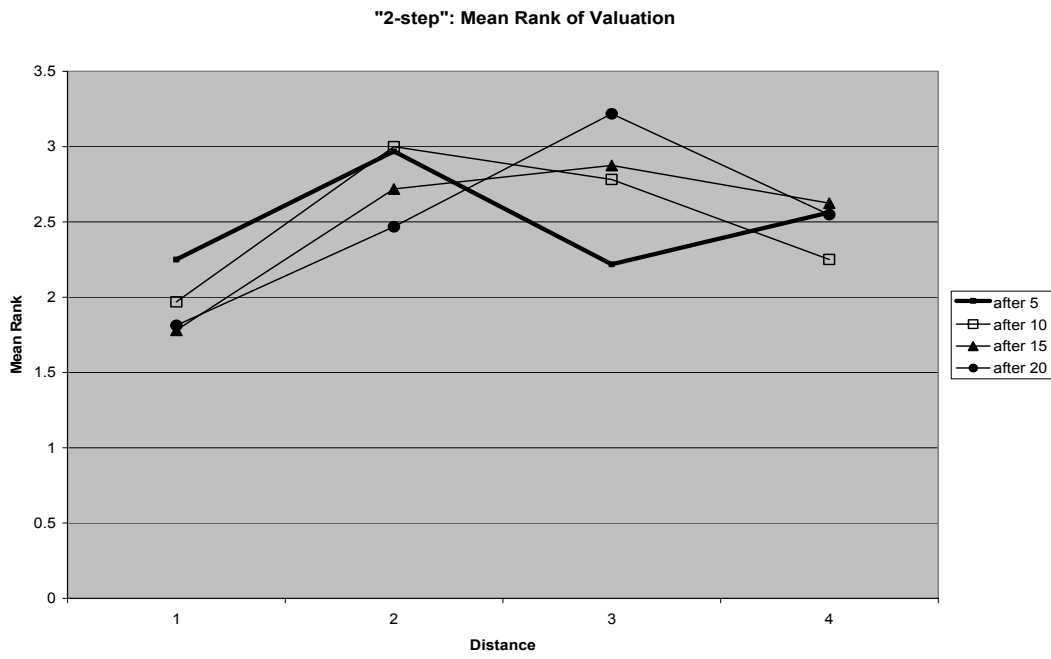


Figure 5b: Mean ranking of valuation for the treatment group '2-step'