Fast Learning in a Simple Probabilistic Visual Environment: A Comparison of ACT-R's Old PG-C and New Reinforcement Learning Algorithms

Franklin P. Tamborello, II (tambo@rice.edu)
Michael D. Byrne (byrne@rice.edu)
Department of Psychology, 6100 S. Main, MS-25
Houston, TX 77005, USA

Abstract

A visual search task used red highlighting to cue the location of the target with varying degrees of probability. The probability that the cue was a valid indicator of target location on any given trial changed during the course of the experiment, and human subjects adapted to this change very rapidly. ACT-R models using the old PG-C and the new reinforcement learning algorithm matched human data from a previous experiment in this paradigm quite well, but only the model that learned by reinforcement mimicked human performance in a new experiment with dynamic highlighting validity.

Introduction

Life is full of environments and tasks people must interact with, and usually they are not perfectly predictable. How do people learn and behave in simple probabilistic environments? Previous research using a simple visual search task included a cue, red highlighting, that had some probability of indicating the location of a target or a distractor, termed "validity" (Fisher & Tan, 1989; Tamborello & Byrne, in press). In brief, the Fisher and Tan task consisted of finding one of four possible targets in a small array of distractors, where the highlighting validity was manipulated as a between-subjects factor. This task is interesting because trials in this task typically take less than one second to complete: will humans be sensitive enough to the probabilistic nature of this rapid environment to adapt their behavior toward efficiency, or will the time-scales involved be too minute for humans to detect? People do appear to optimize a at this level in deterministic environments (Gray & Boehm-Davis, 2000), but it is unclear whether they do so in probabilistic ones.

Tamborello and Byrne found that a cognitive model implemented in the ACT-R cognitive architecture (Anderson et al., 2004) must learn the utility of each and every move of visual attention in the task in order to simulate the differential use of highlighting (termed "sensitivity") to aid visual search. Sensitivity is the response time for trials with invalid highlighting minus the response time for trials with valid highlighting. This quantity is useful as a measure of relative use of highlighting. Tamborello and Byrne's study implemented an ACT-R model that used a learning mechanism, "PG-C" (Anderson et al., 2004), that has since been replaced with a reinforcement learning algorithm (Anderson, 2007; see also ACT-R Research Group, 2007). In brief, ACT-R fires a series of production rules, which are IF-THEN rules stating under what conditions they match, and when they match, what the model does. When multiple production rules match a set of circumstances, they compete. ACT-R resolves the competition by selecting the rule with the highest estimated utility. PG-C estimated a production rule's utility by multiplying the estimated probability (P) of a achieving a goal if that production fires by the value of the goal (G, in seconds), and then subtracting the cost (C, in seconds) of firing that production.

The reinforcement utility learning mechanism now used in ACT-R instead calculates the current production rule's utility as a function of the amount of reward propagated to that production rule. Over many applications the production rule's utility converges on the average amount of reward it receives. Others (e.g., Gray et al., 2006) have claimed that reinforcement learning algorithms work much better than PG-C in certain probabilistic environments, particularly those with costs and rewards at small time scales, such as the Fisher and Tan task. Indeed, Tamborello and Byrne (in press) speculated that this may be why their PG-C model failed to fit their human data very well at low validities. Part of the motivation for this study was to determine whether the reinforcement learning algorithm could do better with low validity highlighting than the PG-C algorithm. Additionally, a new experiment examined human ability to cope with changing environmental probabilities. Could a model built for Tamborello and Byrne's experiment generalize to the new one? Any model that hopes to explain how people behave in probabilistic environments of small time-scales will need to capture major effects from both studies.

The ACT-R Models: Static Validity

Two ACT-R models simulated runs on the current dynamic validity experiment as well as the static validity experiment from Tamborello and Byrne (in press). The previous experiment was identical to the current study's except that highlighting validity remained static throughout the experiment and a wider range of validity conditions were run. In the static validity experiment, highlighting validity was set as a between-subjects factor at increments of 12.5% all the way from 0% to 100%. The same models were run on both the static and dynamic validity experiments.

The models were identical except for which utility learning algorithm they used, PG-C or reinforcement. On any given highlighted trial, the red item was set as the default visual location. For all trials, the default hand location was set to left hand with index finger on "4." With every move of attention, two productions competed:

"attend-red" and "avoid-red." Attend-red requested a move of visual attention to the red item, or else avoid-red requested the location of an unattended black item. If the attended item was red and the target (a "valid" trial), the model could press the appropriate key after a single shift of visual attention. If the attended item was red and a distractor (an "invalid" trial), the reinforcement model propagated a reward of -0.2 (the PG-C model marked a failure) and attended the nearest unattended black item until the target was found. If the model had initially avoided the red item, it could still choose to attend it at any time. This is also a crucial production conflict point because in the case of a standard trial, the models simply moved attention to the nearest unattended item until the target was found. The reinforcement model began a simulation run with a prior utility of 0.01 for the production that would find the red item after the black distractor had been attended. The PG-C model had 75 successes and 25 failures for this same production's priors.

Results and Discussion

Both models fit data from the static highlighting study (the original Tamborello and Byrne experiment) fairly well. The reinforcement model correlated 0.91 (mean deviation 115 ms) with the human data, while the PG-C model correlated 0.89 (mean deviation 110 ms). Figure 1 depicts mean response times (RTs) on valid and invalid trials for the human data, the reinforcement model, and the PG-C model.

There is a particular difficulty for the models in the previous experiment. Assuming subjects really did keep their fingers on the 1–4 number keys, and that the location of the red item was immediately available at trial onset, ACT-R predicts they should take about 300 ms to complete a validly highlighted trial when they initially attend to the highlighted item: 50 ms to decide to attend to the red item, 85 ms to move visual attention, 50 ms to decide to press the appropriate key, then 120 ms to complete the motor movement. Yet in the study reported by Tamborello and Byrne (in press), humans averaged 602 ms to complete highlighted trials when highlighting was 100% valid. That is, given that half of all trials have no highlighting at all, the other half have valid highlighting, and no trials ever have invalid highlighting, people take twice as long on average to complete valid trials as ACT-R's action latencies predict.

To average 600 ms on a trial that should take 300 ms in a best-case scenario, in the wort case subjects must be taking approximately 900 ms to complete the trial, which is about as long as it would take to search the entire five-item display. Were people really avoiding highlighting as often as using it even when it is always valid? Can either the reinforcement or PG-C model capture that effect, or are people perhaps engaging in some action not captured by the models? Humans averaged 602 ms to complete highlighted trials at 100% validity, whereas the reinforcement model averaged 454 ms and the PG-C model averaged 409 ms.

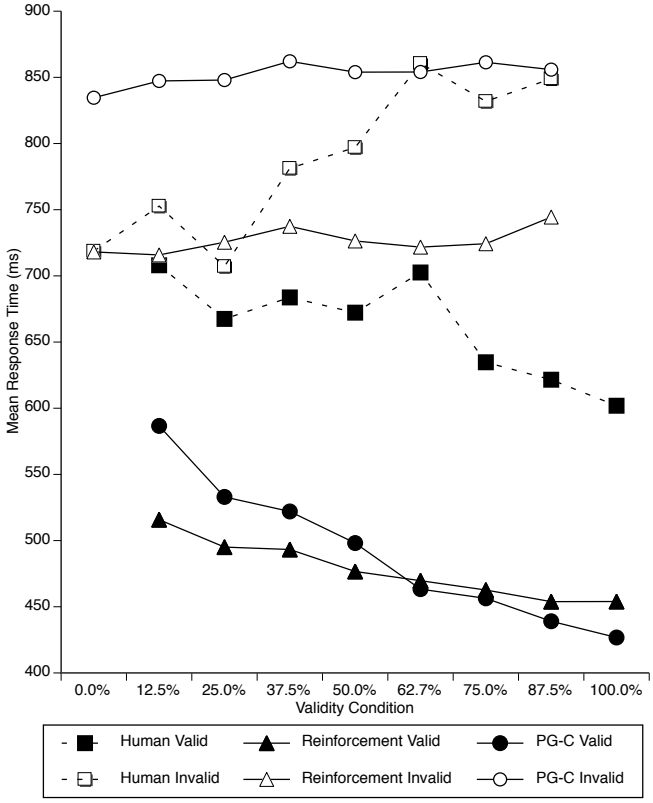

Figure 1. Mean response times of the human data, reinforcement model, and PG-C model for valid and invalid trial types for Tamborello and Byrne (in press) data.

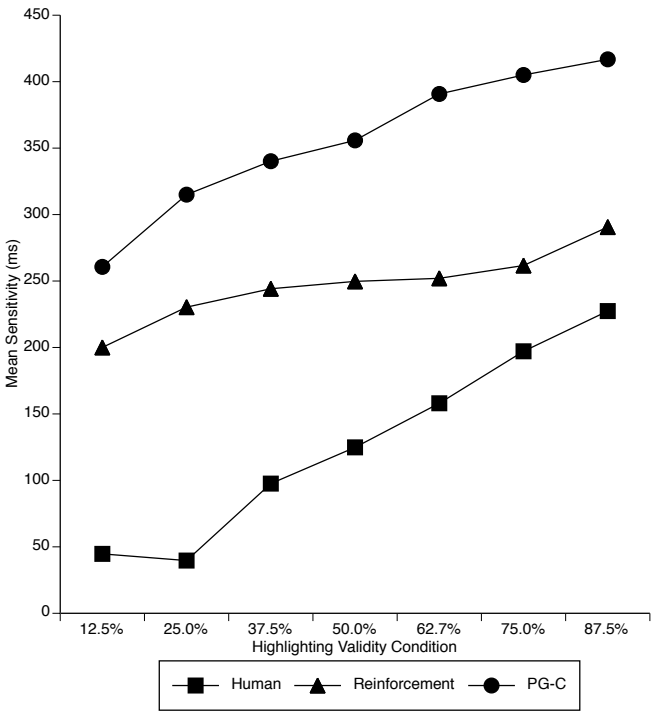

Figure 2. Mean sensitivity per validity condition for humans, the reinforcement model, and the PG-C model for Tamborello and Byrne (in press) data

Figure 2 shows the sensitivity exhibited both by humans and the two models as a function of validity condition. Clearly the reinforcement model does a better job in terms of absolute fit, though the slopes generated by the two models were about equally off (reinforcement model correlation was 0.56; PG-C was 0.53). Interestingly, the PG-C model generated a function which was too steep, and the reinforcement learning model a somewhat too shallow slope.

The Dynamic Validity Experiment

Method

The dynamic validity experiment replicated Tamborello and Byrne's (in press) static validity experiment, except that the validity levels changed during the experiment.

Participants. One hundred nine Rice University undergraduates participated to fulfill experiment participation requirements for their psychology classes.

Design. The experiment incorporated a mixed design utilizing two within-subjects variables, block and trial type (standard, meaning no highlighting; valid, the target was highlighted; and invalid, a distractor was highlighted), and three between-subjects variables: magnitude of validity change, direction of validity change, and change onset timing. Magnitude of validity change refers to by how many percentage points the highlighting validity proportion changed, either 34% or 68%. Direction of validity change refers to whether the highlighting became more valid or less valid. Finally, the experiment was divided into six blocks, affording short rest periods for the subjects between each block. The change in validity occurred at the beginning of either block three (termed "early") or block five ("late"). Thus the total number of between-subjects conditions was eight.

Procedure. Subjects were instructed to place the fingers of their dominant hand on the 1, 2, 3, and 4 keys of the number row at the beginning of the trials and to keep them there throughout the experiment. At the start of a trial, subjects viewed crosshairs for 500 ms at the intended fixation point, in the center of the computer screen. This was then replaced by a horizontal array of five different numerals. The numerals were printed in black 14-point Times New Roman font on a 17-inch CRT computer monitor at a resolution of 1024 by 768 pixels. The highlighting simply used red text. One numeral from the potential target set, {1 2 3 4} was chosen at random, while four distractors from the distractor set {5 6 7 8 9} were also chosen at random. The target and distractors were sorted randomly. The subjects' task was to find the number in the display that was less than five and immediately press the corresponding key on the number row at the top of the keyboard. Subjects were instructed to respond as quickly as possible without making any mistakes. The array disappeared upon the subject's key press, and one second later the next trial began. In the event of an incorrect response, the computer beeped and paused the experiment for two seconds. This time penalty discouraged simple guessing.

Depending upon which condition a subject was assigned to, the initial validity they encountered was 16%, 34%, 68%, or 84%. Blocks consisted of 60 trials, and with the onset of the third or fifth block the validity level changed. Subjects assigned to the 16% initial validity condition then received 84% validity, and vice versa. Similarly, subjects assigned to

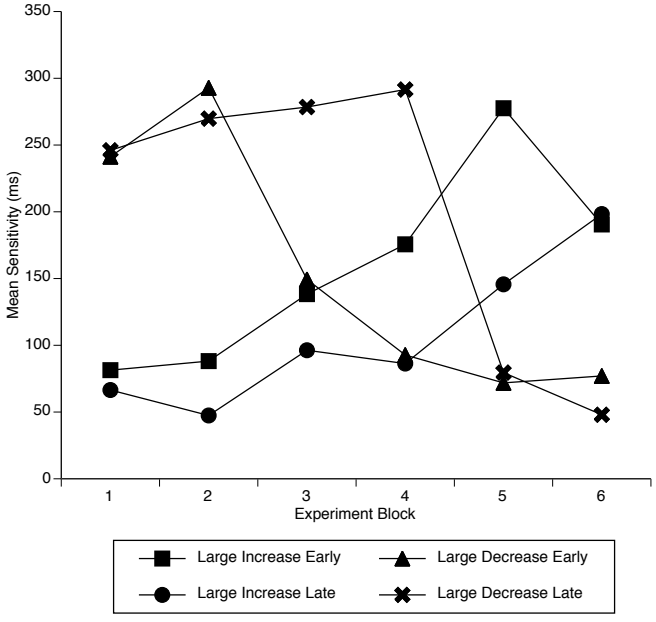

Figure 3. Mean human sensitivities for large change conditions.

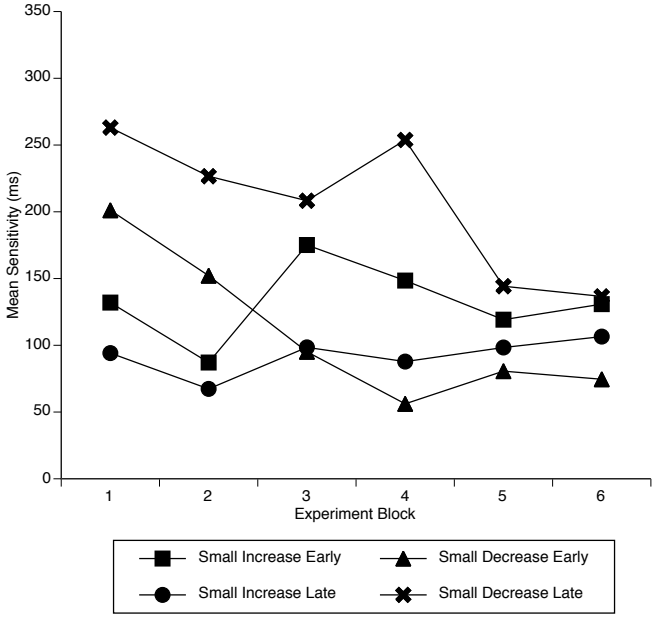

Figure 4. Mean human sensitivities for small change conditions.

the 34% initial validity condition then received 68% validity, and vice versa. The experiment required a little less than 30 minutes for subjects to complete. Instructions did not indicate that the highlighting validity rate would change during the course of the experiment. In many real-world tasks of searching some visual display, such as a web page, the user is not informed a priori about how useful visual cues will be.

Since subjects' sensitivity to changes in validity was assessed by changes in response times, we attempted to avoid confounding with practice effects. Therefore, subjects had a full block of 60 practice trials before beginning the actual experiment. The important task components to practice were searching the field to find the target and pressing the appropriate button in response. Allowing subjects to acclimatize to whatever level of highlighting validity they were assigned to before response times are actually recorded might prove detrimental to the attempt to assess their sensitivity to the highlighting validity. Therefore, practice trials were all of the standard type (no highlighting at all).

Results and Discussion

Outliers were removed prior to statistical analysis. This was done both for single trials and entire subjects. An outlier trial was defined as a trial in which the response time was more than three standard deviations from the subject's overall mean. Those response times were replaced with the subject's mean response time. Each subject's mean response time per condition was similarly screened against the mean response time for all subjects, per condition. Any subject whose mean response time was more than three standard deviations from the mean response time for all subjects in more than one condition was considered an outlier subject. Two such subjects were found, and their data were removed from further analysis. Figure 3 depicts the mean sensitivity for each of the four conditions with large validity proportion changes while Figure 4 depicts those means for the four conditions with small validity proportion changes.

The slopes of the change in sensitivity for the block preceding change onset, the block of the change onset, and the block after the change onset were examined. These three blocks represent prior sensitivity, sensitivity under adjustment, and posterior sensitivity, respectively, and were therefore of most interest for analyzing changing sensitivity in subjects as they adjusted to new highlighting validity proportions. Among the factors size of change, direction of change, and onset of change, only direction had any significant effect on slope, $F(1,98) = 56.13$, $p < 0.01$. There was also a reliable interaction of size with direction, $F(1, 98) = 13.62$, $p < 0.01$. All other F's < 2, p's ≥ 0.17. The direction by size interaction coupled with the main effect of direction means that change direction matters, but it matters more when the change is large than small.

Did subjects in the decreasing validity conditions change their sensitivity faster in response to the changing validity than did subjects in the increasing validity conditions? Presumably subjects would notice a drop in highlighting validity faster than they would notice an increase in validity because of their greater attention paid to a higher prior level of validity. In fact the mean absolute slope for the increase conditions was 56.8 ms per block, and 82.6 for the decrease conditions. A *t*-test of the absolute slope for the two groups failed to yield a reliable difference in degree of sensitivity response to the changing validity in the two groups, $t(104) = 1.92$, $p = 0.58$. The observed effect size in this analysis was

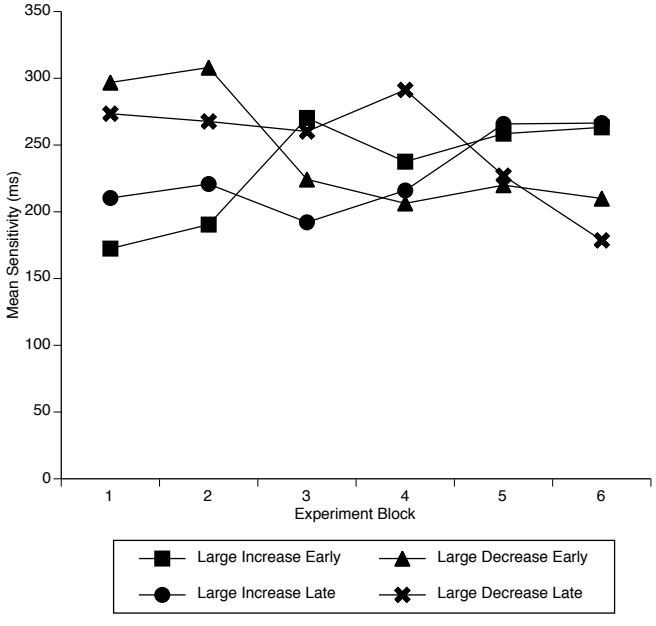

Figure 5. Mean reinforcement model sensitivities for large change conditions.

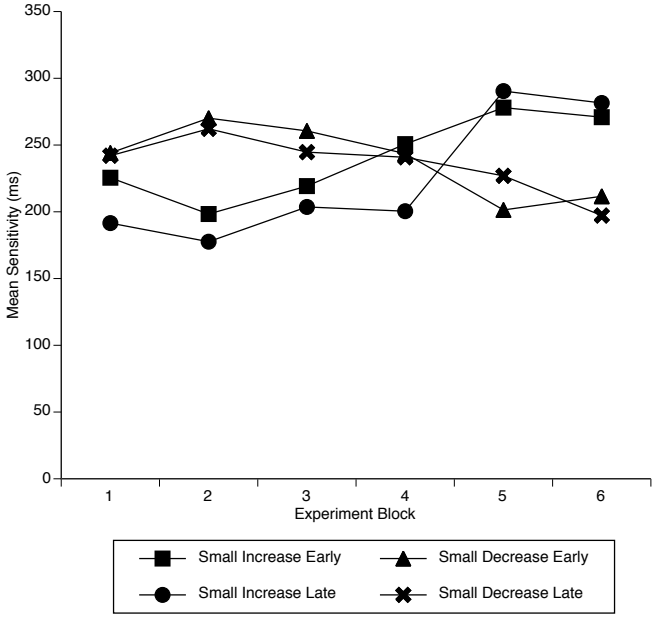

Figure 6. Mean reinforcement model sensitivities for small change conditions.

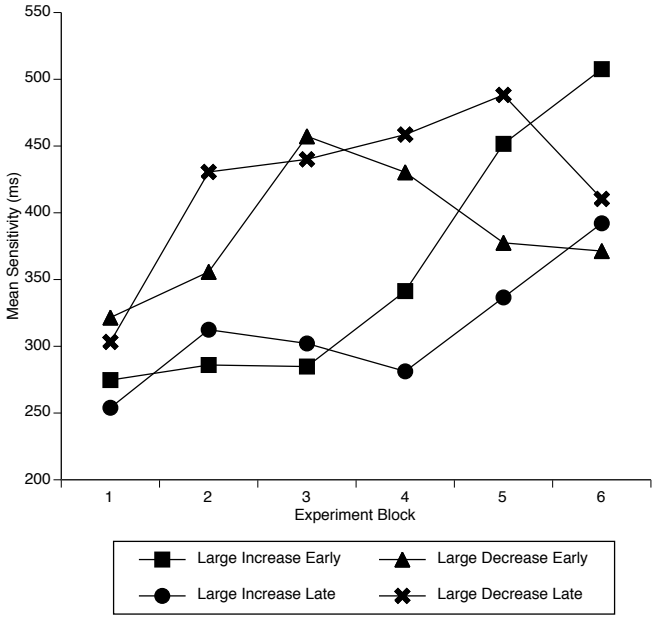

Figure 7. Mean PG-C model sensitivities for large change conditions.

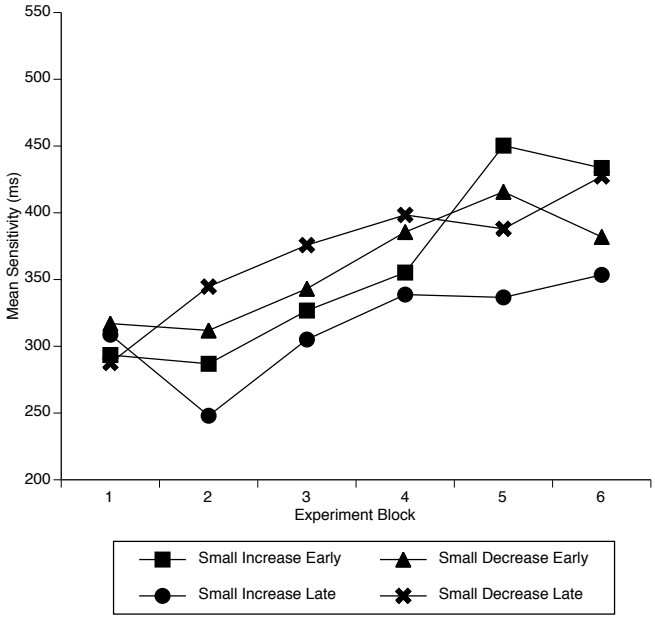

Figure 8. Mean PG-C model sensitivities for small change conditions.

medium-small, Cohen's d = 0.38, and the power to detect an effect of that size was 0.48. The current evidence is therefore inconclusive as to whether the absolute rate of change in sensitivity was different depending upon whether subjects experienced increasing or decreasing validity.

We were also surprised by the lack of reliable effect of late vs. early change; That is, it did not seem matter how much prior experience subjects had with a particular level of validity. Subjects adapted equally well with two additional blocks of experience in a particular validity condition.

The ACT-R Models: Dynamic Validity

As for the dynamic highlighting task, the reinforcement model correlated 0.64 with the human data (mean deviation = 115 ms), while the PG-C model correlated 0.75 (mean deviation = 110 ms). Figures 5 through 8 plot mean sensitivity for the reinforcement and PG-C models. Compare these with Figures 1 and 2. Note how the reinforcement model generally shows the same qualitative trends in its sensitivity functions as do humans. The PG-C model has some hint of those trends, but while the effect size for the large decrease conditions is on the order of 200 ms for the humans and 100 ms for the reinforcement model, it only approximately 50 ms for the PG-C model. Note also how the overall size of the sensitivities keeps increasing throughout the duration of the experiment for the PG-C model (linear F (1, 7) = 29.11, p = 0.001), but not the human data (linear F (1, 7) = 0.88, p = 0.38) nor the reinforcement model (linear F (1, 7) = 0.04, p = 0.84) (Figure 9). The sensitivity function of the reinforcement model generated a slope more closely resembling that of humans (r = 0.86, p = 0.007) than did the PG-C model (r = 0.25, p = 0.55) (Figure 10).

The generally better qualitative fit of the reinforcement model over the PG-C model suggests that learning in a domain like the present study's experiment, a dynamic, probabilistic one of small scale, probably requires a flexible strategy that is more strongly influenced by recent experience than more distantly past experience. One major difference between the standard PG-C algorithm and the reinforcement algorithm is that the reinforcement algorithm uses the last reward propagated to the currently rewarded production to compute the current reward. That last reward, of course, included its previous reward, and so on. However, the PG-C algorithm weighted all past events the same. Lovett (1998) implemented a model of a probabilistic task using a variant of the PG-C utility learning algorithm that incorporated a decay mechanism, though unfortunately this algorithm is computationally expensive. It may be that the need for a decay mechanism to model a probabilistic task indicated the necessity for a fundamental change to ACT-R's utility learning algorithm that would discount distally past experience.

On a side note, a crucial factor in generating reasonable fits to the human data was the degree of prior utility advantage for the production that would seek the red item after a black distractor had been fixated. Models which set this prior too low actually exhibited strong, negative initial sensitivity. We were surprised by how unstable the model's performance was as a function of this single prior utility, which again suggests the critical importance of small time advantages.

Finally, the astute reader will have noticed that the models either attend to the highlighting or avoid it, with no strategy

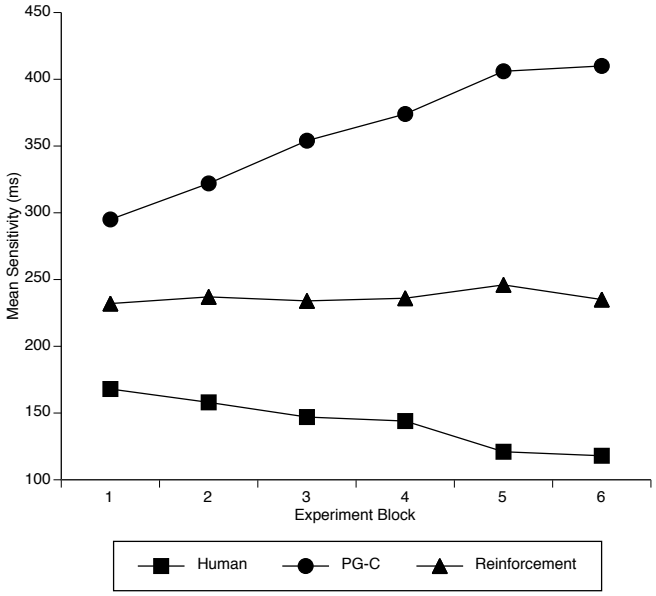

Figure 9. Sensitivity across experiment blocks.

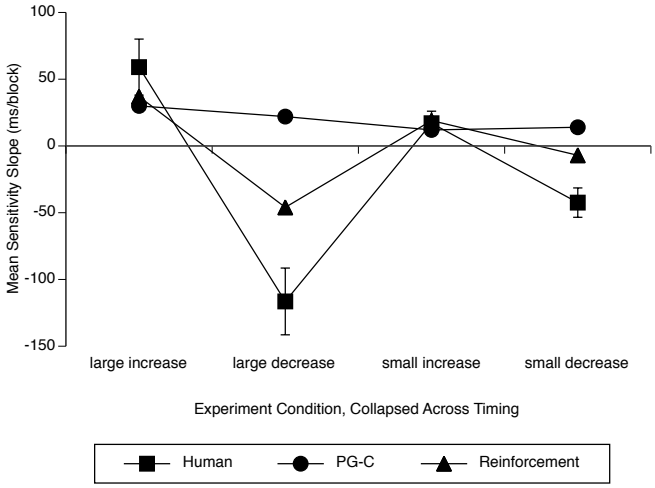

Figure 10. Slope of sensitivity functions, collapsed across timing conditions. Error bars indicate standard error of the mean. Note the interaction of change size and direction in the human and reinforcement model data.

that simply ignores the highlighting. We did this because the highlighting provides information about probable target location as long as validity is not equal to random chance of any one item being the target. When validity is low, one can use highlighting to rule out one search location, and thus by avoiding that location stand to gain approximately 200 ms over a strategy that simply ignores the information provided by highlighting. Gray et al. (2006) demonstrated that people do tend to be efficient in tasks that take place at small time scales. If Gray et al.'s findings generalize to the Fisher & Tan task, then it stands to reason that people will take advantage of information at their disposal for the sake of speed. However, it would still be desirable to actually test this assumption in the future with models that can ignore highlighting rather than or in addition to avoiding it so that such a possibility could be ruled out by data that speak directly to the matter rather than by assumptions based on prior evidence.

References

ACT-R Research Group. (2007). Unit 6: Selecting Productions on the Basis of Their Utilities and Learning these Utilities. Accessed on January 31, 2007 from http://act-r/psy.cmu.edu/

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Quin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036-1060.

Fisher, D.L., & Tan, K.C. (1989). Visual displays: The highlighting paradox. *Human Factors, 31(1)*, 17 – 30.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6*, 322-335.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, 113(3)*, 461 – 482.

Lovett, M. (1998). Choice. In J. R. Anderson & C. Lebiere, *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Tamborello, F. P., II, & Byrne, M. D. (in press). Adaptive but non-optimal visual search behavior in highlighted displays. *Journal of Cognitive Systems Research*.