Adaptive but Non-Optimal Visual Search Behavior in Highlighted Displays

Franklin P. Tamborello, II, Michael D. Byrne {tambo, byrne}@rice.edu
Department of Psychology
Rice University, MS-25
Houston, TX 77005 USA

Abstract

Previous research has suggested that making certain items visually salient, or highlighting, can speed performance in a visual search task. But designers of interfaces cannot always easily anticipate a user's target, and highlighting items other than the target can be associated with performance decrements. An experiment performed suggested that people attend to highlighting less than what an optimal model of visual search in highlighted displays predicts. Users' sensitivity to highlighting's predictiveness depends upon the proportion of trials in which highlighting actually predicts target location. We constructed an ACT-R model that reproduces the major effects of the experiment and which suggests that learning occurs at a very low level of this task.

Introduction

Certain properties of our visual system make it so that certain features, if easily distinguishable from the rest of the visual field, can be especially salient. This visual salience can be harnessed in the form of highlighting to make the visual search component of an information search task more efficient. Yet it is not always easy for the designers of visual interfaces to anticipate the particular target a user may be searching for in a particular context, and it is not clear what the detriments of misleading highlighting may be.

Most investigators agree that certain fundamental features of visual stimuli (such as color, brightness, and movement) are processed in parallel relatively early in the visual pathway of humans (Treisman & Gelade, 1980; Wolfe, 1994). Given a visual search task, the item that can be distinguished by one of those basic features tends to "pop out" from the field of other stimuli. For example, when searching through a field of green T's, the time to find a red T remains roughly the same no matter how many distractor green T's are present. If, however, the target item can only be distinguished by a conjunction of features (such as a certain color and brightness combination), then visual search will be slower and more effortful.

Wolfe's Guided Search 2.0 (1994) theory of visual attention postulates a bottom-up process which filters stimuli through broadly-tuned "categorical" channels. Each of these channels processes a certain dimension of visual stimuli, such as color or orientation. The output of these channels, based on local differences in the stimuli, is then integrated with top-down task demands to compute a feature map for each stimulus dimension. The top-down commands to the feature maps activate locations possessing specific categorical attributes relevant to task demands, such as "activate 'red' objects." The weighted sum of these feature maps forms the activation map, with the weightings being based upon task demands.

Attention deploys limited capacity resources to locations in order of decreasing activation. Visual salience, then, is this combination of bottom-up local difference calculations and top-down, task demand-driven commands. Wolfe actually instantiated his Guided Search 2.0 theory as a computational model of multiple types of visual search tasks. And although Guided Search 2.0 does take into account "top-down" guidance of visual search, it says nothing about learning. That is, how might a human learn about the relative helpfulness of visual cues, and how might that learning be reflected in human performance?

Highlighting by color can be an effective means to harness the computational power of the visual system to aid visual search. Fisher and Tan (1989) performed two experiments assessing the effects of highlighting types and validity on search times. Subjects searched for a target digit in a horizontal array of five digits, one of which was the target. The target was always the digit 1, 2, 3, or 4 (chosen at random), and the distractors were always the digits 5 through 9. Experiment 1 had four highlighting conditions: control (no highlighting), highlighting by color, by reverse video, and by blinking. Additionally, when highlighting was present, the target was highlighted on 50% of the trials, and one of the distractors was highlighted the other 50% of the time. The term "validity" will henceforth be used to describe the proportion of trials on which the highlighting correctly indicated the target. Their experiment 2 was identical, except that highlighting was 100% valid.

Fisher and Tan reported that highlighting by the most effective means, color, did not help participants find the target faster than they found it in control trials when highlighting validity was at 50% (highlighting by other means made subjects slower). Yet when highlighting was 100% valid, subjects did find the target digit 192 ms faster in the color condition than in the control condition. Furthermore, when highlighting validity was 100%, subjects were 90 ms faster on trials with valid color highlighting than they were on trials with valid color highlighting at 50% validity. The authors speculated that the difference in performance occurred because subjects did not always initially attend to the highlighted digit in Experiment 1, but did so in Experiment 2 when they saw that highlighting was more predictive of the items' status as target or distractor. Apparently participants had been making estimates of the relative costs of attending to the highlighting or disregarding it. It was not clear from Fisher and Tan's results what would occur at other levels of validity and so an experiment replicated and extended their results.

One other thing to consider is that people's low-level strategy selection is sensitive to the cost structure of their environment (Gray & Fu, 2004). While Gray and Fu manipulated the cost structure of their subjects' environment by varying the time cost associated with certain actions, we instead manipulated the probability of information being helpful. It is not clear exactly how Gray & Fu's results would generalize to our experiment's probabilistic environment.

The Experiment

The Experiment replicated Fisher and Tan's paradigm, except that only color (red) highlighting was examined and it was examined at additional levels of validity. Fisher and Tan found that subjects were fastest for targets in the middle and roughly equally slow on both ends. This is unsurprising since subjects began each trial with a center fixation. Therefore effects of position were not examined, though target position was randomized. Two experiments were actually run, but will be presented in combination.

Method

Participants One hundred eighty Rice University undergraduates (57% female) participated to fulfill experiment participation requirements for their psychology classes. The participants had a mean age of 19.2 years (1.3). There were 20 subjects per condition.

Design The experiment used a mixed design, with trial type as a within subjects variable: control (no highlighting), valid highlighting, and invalid highlighting. Additionally, subjects were assigned to one of nine validity proportion conditions: 0%, 12.5%, 25%, 37.5%, 50%, 67.5%, 75%, 82.5%, or 100% highlighting validity. Half of all trials each subject received was of the control type. A person in the 75% highlighting validity percentage condition, for example, would receive half control trials and half highlighted trials. Of the highlighted trials, 75% of those would have valid highlighting, 25% would have invalid highlighting.

Procedure The subjects' task was to, as quickly as possible without making any mistakes, 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. At the start of the trial, subjects viewed crosshairs for 500 ms at the intended fixation point in the center of the computer screen. They subsequently viewed a horizontal array of five different numerals. The numerals were printed in black (red for highlighted items) 14-point Times New Roman font on a 17-inch CRT computer monitor at a resolution of 1024 by 768 pixels. At a typical viewing distance of approximately 60 cm, the entire array subtended a visual angle of approximately 8° from left side of the left-most digit to the left side of the right-most digit. There was approximately 2° of visual angle from the left side of one digit to the left side of an adjacent digit.

One digit 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. One target was present during every trial. The target and distractors were sorted randomly. The array would disappear upon the subject's key press, and one second later the next trial would begin. In the event of an incorrect response, the computer beeped and paused the experiment for two seconds. This time penalty discouraged simple guessing. There were six blocks of 64 trials each.

Results and Discussion

Figure 1 summarizes the means for each validity percentage condition by trial type. There was an interaction of trial type and validity percentage condition such that as validity increased, subjects became faster on valid trials and slower on invalid trials, *F*(12, 258) = 16.22, $p < 0.001$. But there was no main effect of validity condition on response times (RT) for the control trial type, *F*(6, 129) = 1.14, $p = 0.341$.

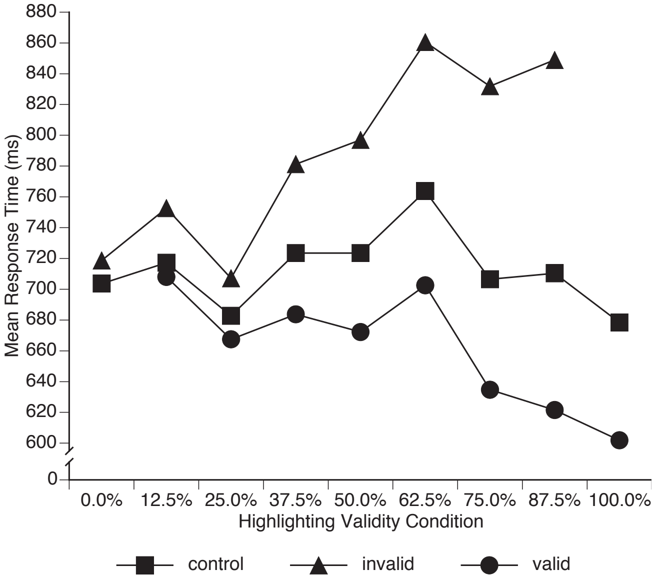

Figure 1. Mean RT as a function of validity percentage condition.

Collapsing across validity percentage conditions, subjects responded to valid trials on average 51 ms faster than control trials, for which they were 75 ms faster than invalid trials (linear *F*(1, 129) = 418.32, $p < 0.001$). Another way to look at these data is to consider how sensitive subjects were to the highlighting. If subjects always attended the highlighted item first, then they should show fast RTs to valid trials and slow RTs to invalid trials. If they ignored the highlighting, then these two RTs should be about the same. Thus "sensitivity" was defined as the RT for invalid trials minus the RT for valid trials. Subjects did indeed show higher sensitivity at higher levels of validity. Note in Figure 1 how the lines for valid and invalid trials diverge going from the 0% validity percentage condition to the 100% validity percentage condition. Furthermore, participants sensitized within the first block.

Figure 2 plots the effect of validity percentage condition on sensitivity in block 1 only. Note how the effect of validity on sensitivity is ordinal: sensitivity increases as validity increases. Within block 1, subjects in the 87.5% validity percentage condition were 201 ms more sensitive to trial validity than were subjects in the 12.5% condition, an over four-fold difference. ANOVA with linear contrast indicated an increasing effect of validity percentage condition such that subjects were more sensitive to trial type the more often they encountered validly highlighted trials, *F*(1, 129) = 45.362, $p < 0.001$. Furthermore, this trend held for all trial blocks, linear *F*(1, 129) = 54.06, $p < 0.001$, though sensitivity did decline overall from block one to block six, linear *F*(1, 129) = 12.90, $p < 0.001$.

The analysis of subjects' sensitivity to highlighting indicates that subjects learned rapidly whether or not they could take advantage of highlighting in their visual search. The more valid the highlighting was, the more subjects used

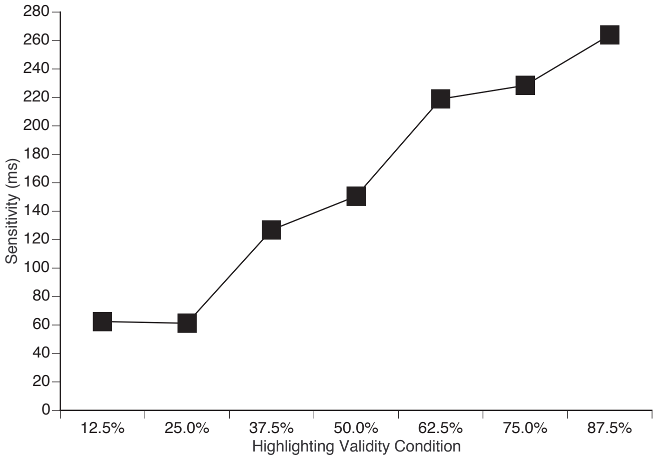

Figure 2. Sensitivity in block 1 as a function of validity percentage condition.

it and the faster their RTs for valid trials became–and this was true even in the first block. However, as validity increased, so did the cost of using highlighting on the occasions that it was invalid.

The Algebraic Model of Optimality

Fisher and colleagues (1989) constructed an algebraic model of visual search in an attempt to determine average search time for a target in displays with or without highlighted options. The model computes expected search times given total set size, highlighted subset size, probability of the target being in the highlighted or unhighlighted subsets, and probability that the subject searches first in the highlighted or unhighlighted subset. Readers are encouraged to refer to Fisher et al.'s article for a detailed description of the model. The authors concluded that only a model that captures validity probability can account for observed variance of RT as a function of highlighting validity probability. When the display consists of a set of discrete options, as in our experiment, then the average display search time is a probability mixture of the time to find the target when the highlighted options are searched first and the time to find the target when the unhighlighted options are searched first. Their model assumes users exhaustively search through one subset before searching the other.

Our instantiation of the Fisher et al. model assumes 50 ms to decide to attend highlighting at the beginning of a trial, 85 ms to shift visual attention, 100 ms to decide to attend a new location, and 150 ms to issue a response. These latencies are based on ACT-R's predictions of how long these actions should take. Figure 3 plots the predicted mean RTs of Fisher et al.'s optimal model of visual search in the 5 x 1 array paradigm for trials with valid and invalid highlighting. Also shown is data collected from human subjects. Note that the optimal model predicts a highlighting sensitivity in excess of 300 ms for even the low validity percentage conditions, whereas the humans showed sensitivity in these conditions more on the order of 40 or 50 ms. Note also how the predicted RTs for valid trials are approximately 350 ms less than those obtained. Predicted RTs for invalid trials also were reliably less than those obtained from humans, $t(14) = 2.409$, $p = 0.030$, and the human RTs increased at a faster rate as a function of increasing validity level, ANOVA with linear contrast on highlighting validity $F(1,14) = 5.803$, $p = 0.030$.

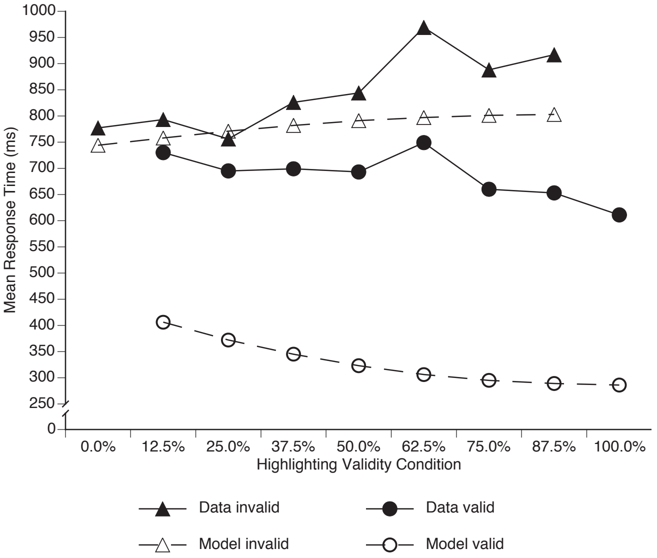

Figure 3. Predicted RTs of Fisher et al.'s optimal model and RTs from human data.

The optimal model computes P_H, the probability that the subject first examines the highlighted subset. Table 1 lists P_H for each level of highlighting validity. When both the total number of items and number of highlighted items is low, as it is in this experiment, P_H will always be high because the random chance of finding the target in any given position is high (20%). The high P_H is what drives the model to predict such fast times for valid trials, and it may be that subjects are not as fast in valid trials because they are not so likely to search the highlighted item first.

Finally, Fisher et al.'s algebraic model does not take into consideration the possibility that subjects may attend to some subset of the unhighlighted options, then attend to the highlighted option, then if necessary, finish searching the unhighlighted options. As shall be shown with the ACT-R models, the ability to attend to the highlighting at any time may be important for replicating the human data.

What contributes to the adaptability of human visual search behavior in an environment where cues have some probability of being helpful or hindering? Given a fairly simple paradigm like the digit search in a 5 x 1 array used by Fisher and Tan,

Table 1: Values for P_H given highlighting validity.

Validity	P_H
0.0%	0.785
12.5%	0.837
25.0%	0.882
37.5%	0.919
50.0%	0.949
62.5%	0.971
75.0%	0.986
87.5%	0.994
100.0%	0.998

there are a couple of decisions that could be made by the user at a variety of different times throughout the task. Users can potentially elect to attend or avoid the highlighting each time they shift their visual attention, though they might not actually make that decision that often. Then there are issues of self-evaluation of performance: do users evaluate their own performance at the level of whole trials or some smaller level, such as individual shifts of attention? Fisher et al.'s algebraic model has no answer to offer for these questions.

An ACT-R model can shed some light on the above questions, since it must be explicitly programmed to attend or avoid highlighting either with every shift of attention or only the first one. It could also evaluate its own performance at either a macro level (i.e., a full trial) or micro level (i.e., each shift of attention). Two ACT-R models were constructed to compare to Fisher et al.'s algebraic model and to make inferences about people's search behavior in reference to the micro versus macro issues discussed above.

The ACT-R Models

The ACT-R (Anderson et al., 2004) cognitive architecture was used to construct two models of this task. The two models, macro-level and micro-level, were identical except in the level at which they learn: whole trials or individual shifts of attention, respectively. On any given trial with highlighting present, the models selected and fired one of two productions which caused them to attend to (”attend-red”) or avoid the red item (”avoid-red”). At the start of each highlighted trial, both productions match the contents of the buffers, which include a buffer-stuffed red item. “Attend-red” made the models move visual attention to the red item. “Avoid-red” made them seek an unattended black item. If the currently attended item was a target, the models output a press of the appropriate key.

If the item was a distractor, a production, “highlighted-distractor”, fired. This production was marked as a failure in the micro-level model, but it was left unmarked for the macro-level model and is the critical difference between the two models. If the red item was still unattended, then the models still had the decision to make as to whether to attend to or avoid the highlighting (“avoided-red-distractor-find-red” and “avoided-red-distractor-find-black”, respectively). In the case of control trials, the models simply fixated unattended items until they found the target.

Production priors were set as in Table 2. The priors for “avoid-red” were set equal to the random chance that the target was in either the highlighted or unhighlighted subset, 20% chance for the highlighted group since one item would be highlighted out of a total of five items in the display. Priors for “avoided-red-distractor-find-red” and “avoided-red-distractor-find-black” were set equal to the random chance that the target would be in either highlighted or unhighlighted subset, given that one unhighlighted item was already examined and found to be a distractor. The probabilities for successes and failures for “attend-red” were set following the same rule, thus 20 successes and 80 failures. But it was found that the model did not attend the highlighted item often enough to generate a similar RT pattern to the human data. This could imply that people may be inclined to check highlighted subsets at rates greater than chance.

Table 2: Production priors for the ACT-R model.

Production	Successes	Failures	Initial Utility
attend-red	50	50	9.95
avoid-red	80	20	15.90
avoided-red-distractor-find-red	25	75	4.90
avoided-red-distractor-find-black	75	25	14.90

$$U_i = P_i G - C_i \tag{1}$$

$$P = \frac{Successes}{Successes + Failures} \tag{2}$$

One consequence of this set of priors is that the model will be slightly biased toward examining an unhighlighted item first, but not as biased as if it were simply ignoring highlighting altogether and checking each item randomly. ACT-R computes the utility of every production *i* using Equation 1, where P_i is an estimate of the probability that if production *i* is chosen the current goal will be achieved, *G* is the value of that current goal, and C_i is an estimate of the cost (typically measured in time) to achieve that goal (Anderson et al., 2004). Equation 2 computes *P* and C_i is the simulated time elapsed between firing production *i* and firing another production that is marked as a success. Thus “attend-red”'s utility started at .5 * 20 - 0.05 = 9.95 versus 15.90 for “avoid-red.”

This seems valid as a set of starting assumptions about the expectations about highlighting subjects brought with them to the experiment given a lack of measurement of the match between subjects' information search goals over the course of their lives and the highlighted items they have encountered. The priors were scaled to values between 10 and 100 such that the model would start a run with some biases in the productions it chose, but not so biased as to be inflexible to learning new utilities over the course of the 384 trials.

Two ACT-R models were used to explore the question: When is an action a success or a failure? If learning occurs at the macro (trial) level, all trials are successful because the target is always eventually found. Learning, then, occurs only as a result of a trial taking relatively more or less time to complete. But if learning takes place at a micro level (individual attention shifts), then all attention shifts toward distractors are failures and only the attention shift toward the target is a success. Learning only by time elapsed might not be enough feedback to change behavior. If so, the model may need the failures encountered when attention shifts to distractors are explicitly tagged as failures.

At trial onset (highlighted trials), the models had two competing productions: “attend-red” and “avoid-red”. Subsequent shifts of attention within the trial had similar productions that competed to make the model either attend to the highlighted item if it had not already, or else avoid the highlighted item. The models were run with the expected gain noise parameter set to 3 because it resulted in the best

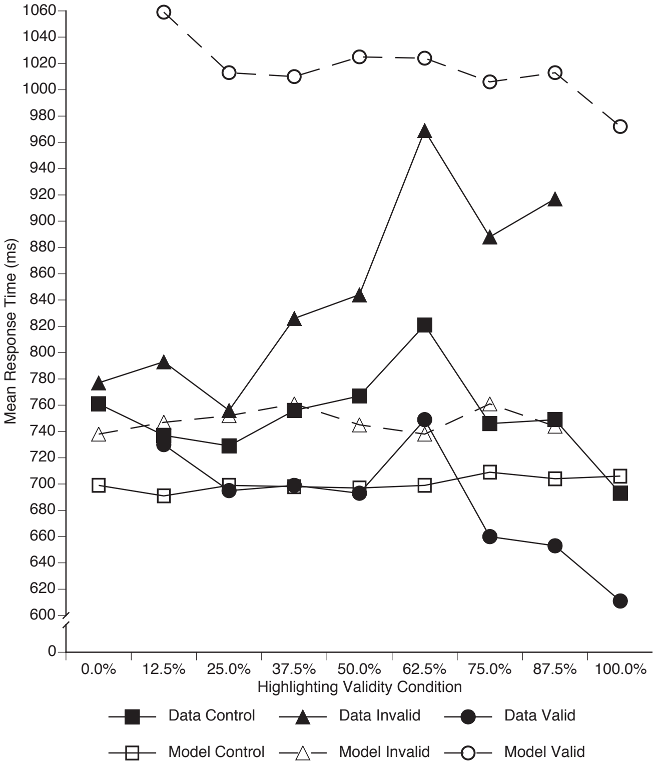

Figure 4. Human data versus ACT-R macro-level model data.

fit to the human data. The utility threshold was set to –100 to ensure a matching production always fired.

Figure 4 plots the mean RTs for human and the macro-level model's data, per trial type, per highlighting validity percentage condition. Note that unlike the human data, the macro-level model does not change sensitivity across highlighting validity percentage conditions. Note that mean RTs for valid trials are almost all in excess of 1,000 ms while mean RTs for invalid trials are all approximately 740 ms. The macro-level model thus has a comparatively large and inverted sensitivity function with respect to the human data.

By contrast, the micro-level model unequivocally produced the better fit to the human data, r2 = 0.964, mean deviation = 67 ms (figure 5). As highlighting validity percentage increased, valid trials became faster and invalid trials became slower, as in the human data. The one major flaw in the model is its poor prediction of RTs for valid and invalid trials in low validity percentage conditions. Because it is too slow on such trials, the model shows too much sensitivity in low validity percentage conditions. More particularly, the model predicts RTs that are too slow for invalid trials at low validity levels, and does not increase enough as validity increases.

While we have determined that the micro-level model actually does avoid red items all the way through invalid trials in the low validity percentage conditions, it is not readily apparent why the model generates such slow RTs for these trials. We hope that detailed examination of proportions of firings of "attend-red," "avoid-red," and their subsequent shift of attention counterparts will yield clues as to what the model is actually doing in these trials. Despite this flaw, the model's prediction of the effect of validity on sensitivity does closely mimic that of the human data (Figure 6), r2 = .916. The high correlation between model sensitivity and human sensitivity may be misleading because the model does fail to accurately predict RTs for low validity conditions and is generally 50 ms too fast for control trials. Despite this the model succeeds very well in capturing the effect of validity percentage condition on sensitivity. The micro-level ACT-R model's successful capture of the sensitivity effect is important because it allows the model to display the same qualitative trends as humans: it generally gets faster for highlighted trials as validity percentage increases because it learns to attend to the highlighting more. However as validity percentage increases, so does mean RT for invalid trials because the model more often attends to highlighting, including the occasions when it is still misleading.

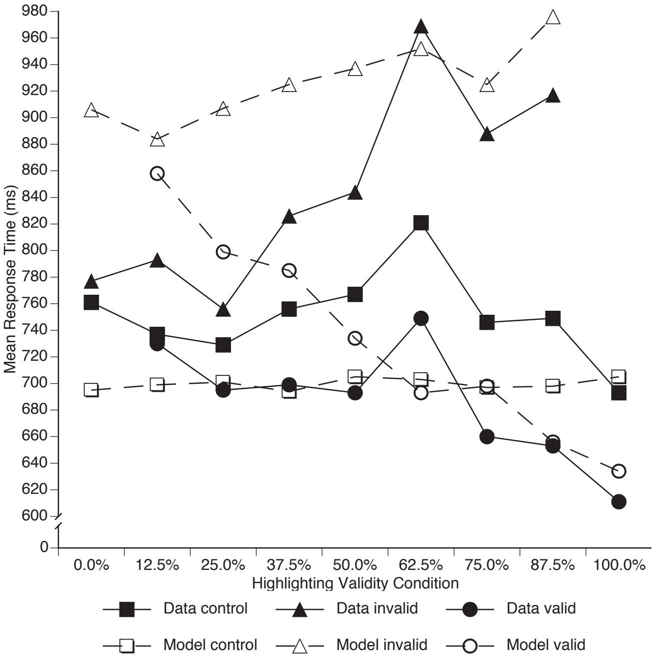

Figure 5. Human data versus ACT-R micro-level model data.

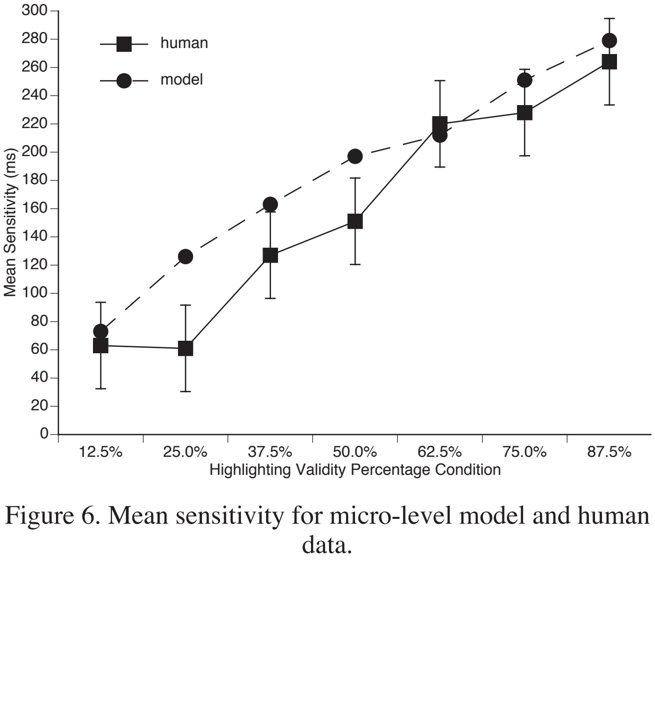

Figure 6. Mean sensitivity for micro-level model and human data.

General Discussion

Fisher et al.'s (1989) optimal model of visual search of highlighted displays assumes that people do not always check the highlighted subset first, but when the total search space is small this model seems to overestimate the degree to which people do check that group first. Why do people fail to behave optimally, to always attend the highlighted item first when validity exceeds the probability of finding the target by random chance?

Even in very simple experiments employing a probabilistic environment that rewards optimizing behavior, people tend to match the probabilities of their responses to outcomes rather than optimize their behavior (Shanks, Tunney, & McCarthy, 2002). In this task those probabilities matter for costs and rewards because a shift of attention to the wrong item translates into wasted time. There is some evidence that ACT-R's current production utility learning algorithm does not appropriately capture the dynamics of the environment with respect to cost and reward (Gray et al., in press), and that may be why the model fits poorly in the low highlighting validity percentage conditions.

The ACT-R models reveal that people seem to make decisions about attending and avoiding highlighting at the level of individual shifts of attention, rather than at the level of an entire trial as in the Fisher et al. model. To do that people are going to need to evaluate their own performance at that microscopic level. There is evidence from other similar domains that events on such a small time scale do have behavioral consequences (Gray & Fu, 2004). To some degree, people are learning the probabilities with which highlighting leads them to the target and they are using that knowledge to exercise control over where they attend. This learning and deciding where to attend at the micro level may be people's best way to take advantage of the pop-out effect in an environment where what is visually salient may not always be what they are looking for.

References

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., Coury, B. G., Tengs, T. O., & Duffy, S. A. (1989). Minimizing the time to search visual displays: The role of highlighting. Human Factors, 31(2). 167 – 182.

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

Gray, W. D., & Fu, W. T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. Cognitive Science, 28. 359 – 382.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (in press). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. Psychological Review.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. Journal of Behavioral Decision Making, 15. 233 – 250.

Treisman, A.M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12. 97 – 136.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. Psychonomic Bulletin & Review, 1(2). 202 – 238.