Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 8 (2007) 182–191

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Adaptive but non-optimal visual search behavior with highlighted displays ☆

Action editor: Andrea Stocco

Franklin P. Tamborello II *, Michael D. Byrne

Department of Psychology, MS-25, Rice University Houston, TX, USA

Received 21 February 2007; accepted 10 May 2007
Available online 3 July 2007

Abstract

Previous research [Fisher, D. L., & Tan, K. C. (1989). Visual displays: The highlighting paradox. *Human Factors, 31*(1), 17–30] suggested that making certain items visually salient, or highlighting, can speed performance in visual search tasks. But interface designers cannot always anticipate users' intended targets, and highlighting non-target items can lead to performance decrements. An experiment presented suggests that people attend to highlighting less than what an algebraic visual search model of highlighted displays [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] predicts. Users adjust their visual search strategies by probability-matching to their visual environment. An ACT-R [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] model reproduced the major effects of the experiment and suggests that learning in this task occurs at very small cognitive and time scales.

Keywords: ACT-R; Learning; Probabilistic environment; Visual search

1. 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 objects, the time to find a red object remains roughly the same no matter how many distractor green objects are present. If, however, the target item can only be distinguished by a conjunction of features (such as a certain color and shape 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

☆ This article represents an expansion of an article we presented at the Seventh International Conference on Cognitive Modeling in Trieste, Italy, April 5–8, 2006. The authors thank their reviewers for the helpful comments they provided. It should be noted that Figs. 4 and 5 of the ICCM publication contain inaccuracies in the human data depicted, and the same inaccurate human data had been used to evaluate fit of the micro-level ACT-R model. The authors accept full responsibility for those inaccuracies and have corrected them in this manuscript.

* Corresponding author. Tel.: +1 713 348 2141.
E-mail address: tambo@rice.edu (F.P. Tamborello II).

1389-0417/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2007.05.003

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. Although Guided Search 2.0 does take into account "top-down" guidance of visual search, it says nothing about learning in visual search tasks. 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 on each trial), and the distractors were always selected from 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 when highlighting validity was at 50% highlighting by the most effective means, color, did not help participants find the target faster than they found it in control trials (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 our 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; Gray, Sims, Fu, & Schoelles, 2006). 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.

2. The experiment

The experiment replicated Fisher and Tan's paradigm, except that only color (red) highlighting was examined and it was examined at multiple 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.

2.1. Method

2.1.1. Participants

One hundred eighty Rice University undergraduates (57% female) participated for course credit. The participants had a mean age of 19.2 years (standard deviation 1.3). There were 20 subjects per condition. Participants were not screened for color vision deficiency. Although some 8% of the male population and 0.5% of the female population may have some form of color vision anomaly (Wolfe et al., 2006), even if a handful of subjects responded in no way to highlighting it would only mean that their data would tend to dilute effects of highlighting on normal color visioned individuals. The rather robust effects of highlighting reported in the results section, particularly for main effect of trial type, obviate any cause for concern from this issue.

2.1.2. 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%, 62.5%, 75%, 87.5%, or 100% highlighting validity. Control trials comprised half of all trials received by each subject. A person in the 75% highlighting validity 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. All three trial types were distributed randomly within each trial block, and all trial blocks had all trial types in the same proportions as dictated by experimental condition.

2.1.3. 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. The advantage to not telling subjects which target would appear is that it forced subjects to really search the display. Any errors could easily be identified and excluded from analysis as such trials did not signify a successful search of the display. 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-in. 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.

The experiment consisted of six blocks of 64 trials, and at the conclusion of each block participants were encouraged to rest for perhaps a minute or two before continuing to the next block. Most participants completed the experiment within 35 min. For each trial 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 without replacement. 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.

2.2. Results and discussion

Subjects erred on 3140 of a total of 69,120 trials administered, or fewer than 5% of total trials. Analyses excluded error data as errors on such a simple task do not reflect a successful search of the display. Outliers were removed prior to statistical analysis, and 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 total mean response time for all subjects, per condition. Any subject whose mean response time was more than three standard deviations from the total mean response time for all subjects in more than one condition was considered an outlier subject. Four such subjects were found, and their data were removed from further analysis.

Fig. 1 summarizes the means for each validity percentage condition by trial type. There was an interaction of trial type and validity 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, $F(6,129) = 1.14$, $p = 0.341$.

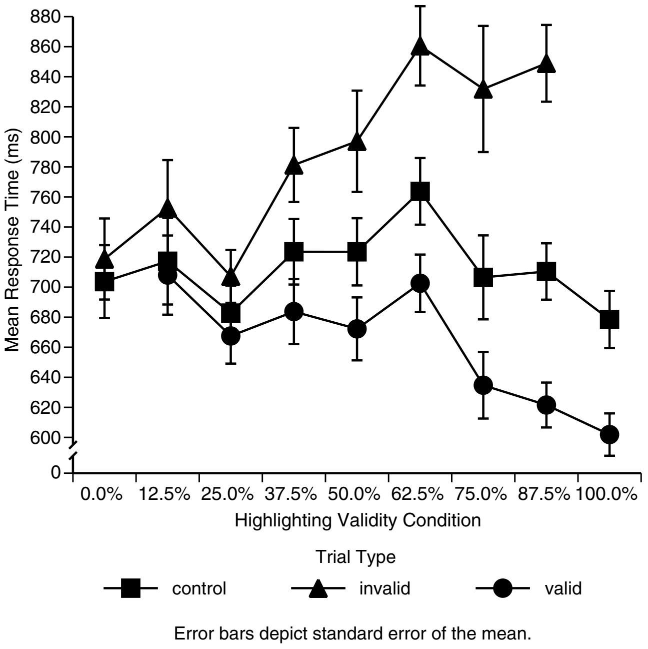

Fig. 1. Mean RT as a function of validity condition.

Collapsing across validity 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 Fig. 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' RTs showed sensitivity to trial type within the first block, indicating that they immediately utilized the highlighting in their visual search.

Fig. 2 plots the effect of validity condition on sensitivity in block 1 only. Note how the effect of validity on sensitivity is nearly linear: sensitivity increases as validity increases. Within block 1, subjects in the 87.5% validity condition were 201 ms more sensitive to trial validity than were subjects in the 12.5% condition, an over fourfold difference. ANOVA with linear contrast indicated an increasing effect of validity 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 $F(1,129) = 12.90$, $p < 0.001$ (see Fig. 3). That effect is driven by subjects getting faster

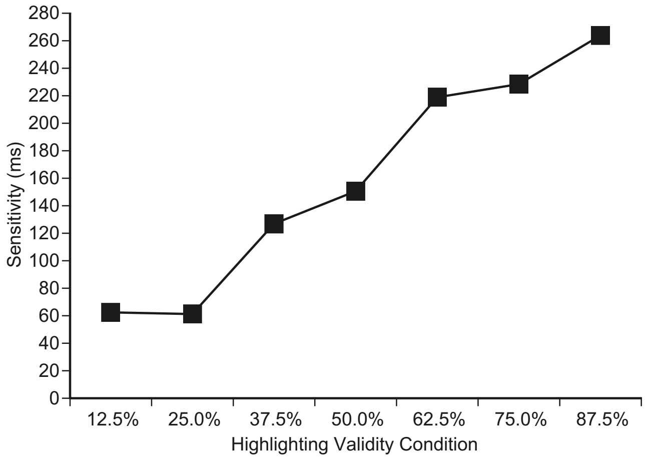

Fig. 2. Sensitivity in block 1 as a function of validity percentage condition.

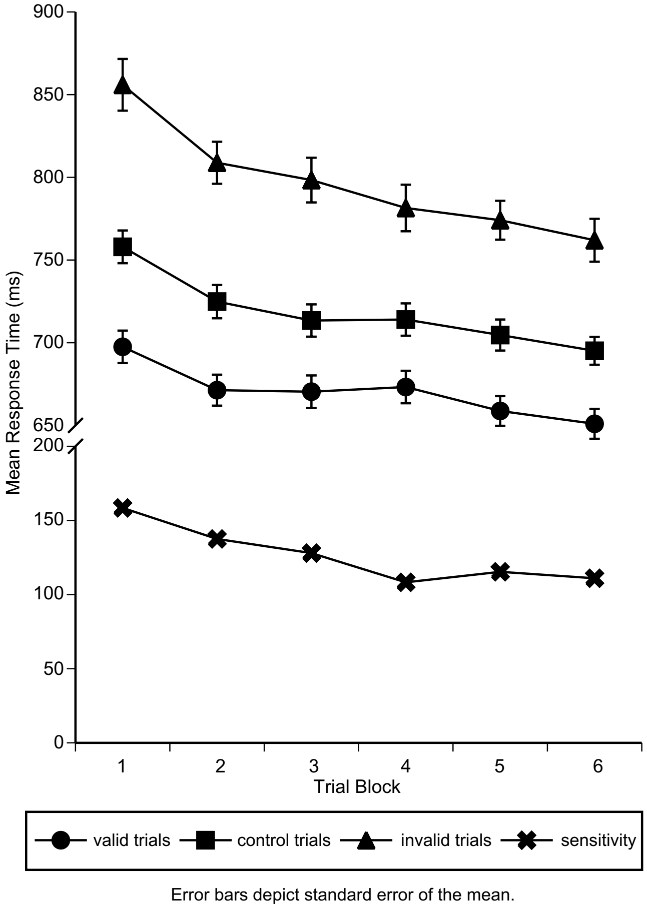

Fig. 3. Mean response time for each trial type and mean sensitivity, plotted by block.

on invalid trials (94 ms) to a greater degree than they get faster on valid trials (47 ms), linear by linear interaction effect $F(1,129) = 19.70$, $p < 0.001$. There is some, albeit unreliable, suggestion that this effect varied by highlighting validity condition, interaction $F(30,645) = 1.46$, $p = 0.056$. Even if subjects did improve their performance at different rates in different validity conditions, it is not clear what this effect may signify.

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 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.

2.3. The algebraic model

Fisher, Coury, Tengs, and Duffy (1989) constructed an algebraic model of this type of visual search in an attempt to determine average search time for a target in displays with or without highlighted items. 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 (Eq. (1)). The authors concluded that a model accounting for the observed variance of RT must capture highlighting validity probability. When the display consists of a set of discrete options, as in the Fisher and Tan (1989) paradigm, 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. Fisher et al.'s model assumes that users exhaustively search through one subset before searching the other. This is an important point that will be explored further in the ACT-R models presented below.

$$E_{\mathrm{H}}[T] = e\left\{[(h+1)/2]\left[\sum_{i=1}^{h} P(O_i)\right] + [(n+h+1)/2] \times \left[\sum_{i=h+1}^{n} P(O_i)\right] + nP(O_{n+1})\right\} + r \qquad (1)$$

Eq. (1) states that the expected time to find the target given that the subject searches first the highlighted set, $E_{\mathrm{H}}[T]$, is the sum of four quantities. The scaling quantity, e, is the expected encoding time for each item. The first expression inside the curly braces is the average number of highlighted items searched through times the probability that the target is present and highlighted. Note that in the present study's experiment the probability that the target is present in any trial is 1. The second expression inside the curly braces is average number of unhighlighted items searched times the probability that the target is present but not highlighted. The third expression inside the curly braces is the product of the conditional expected search time given that the target is absent times the probability that the target is absent. The fourth quantity is the expected response time, in other words, the time required for the subject to decide to press a certain button and then to execute that motor movement. The expected response time when the unhighlighted options are searched first is similarly computed, as shown in Eq. (2).

$$E_{\mathrm{H'}}[T] = e\left\{[(2n-h+1)/2]\left[\sum_{i=1}^{h}P(O_i)\right] + [(n-h+1)/2] \times \left[\sum_{i=h+1}^{n}P(O_i)\right] + nP(O_{n+1})\right\} + r \qquad (2)$$

Assuming the display consists of a set of discrete options, as it does in the above experiment, then the average display search time can be derived. The total expected time to find the target in the display is a sum of two expressions (Eq. (3)). The first is the probability of searching first through the highlighted subset times the expected time to find the target when subjects search first through the set of highlighted options. The second is the probability of searching first through the unhighlighted subset times the expected time to find the target when subjects search first through the set of unhighlighted options. In other words, it is the sum of two expected values: the expected time to find the target when the highlighted options are searched first and the expected time to find the target when the unhighlighted options are searched first.

$$E[T] = p_{\mathrm{H}}E_{\mathrm{H}}[T] + (1-p_{\mathrm{H}})E_{\mathrm{H'}}[T] \qquad (3)$$

Fisher et al. (1989) assume that p_{H} (probability that subjects search first through the highlighted subset) should increase as the average time it takes to locate the target by attending first to the unhighlighted options grows proportionately larger than the average time it takes to locate the target by attending first to the highlighted options (i.e., as the ratio $E_{\mathrm{H'}}[T]/E_{\mathrm{H}}[T]$ increases). Eq. (4) is the function they offer that satisfies that constraint.

$$p_{\mathrm{H}} = 1 - \mathrm{e}^{\{2.0(E_{\mathrm{H'}}[T]/E_{\mathrm{H}}[T])\}} \qquad (4)$$

Fisher et al. (1989) claimed their model fit data from their experiment with a 36-item display when 1, 3, 6, or 12 items were highlighted, but people are often faced with simpler displays. How well would the Fisher et al. model fit data from the Fisher and Tan (1989) experiment, which used a display of only 5 items?

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 the ACT-R durations for these operations, which are reasonably well-established parameters. Readers are invited to consult Anderson et al. (2004) for a summary of ACT-R and the generalized theory of human cognition it represents. Fig. 4 plots the predicted mean RTs of Fisher et al.'s algebraic model of visual search in the 5 × 1 array paradigm for trials with valid and invalid highlighting. Also shown is data collected from human subjects. Note that the algebraic 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 $F(1,14) = 5.803$, $p = 0.030$.

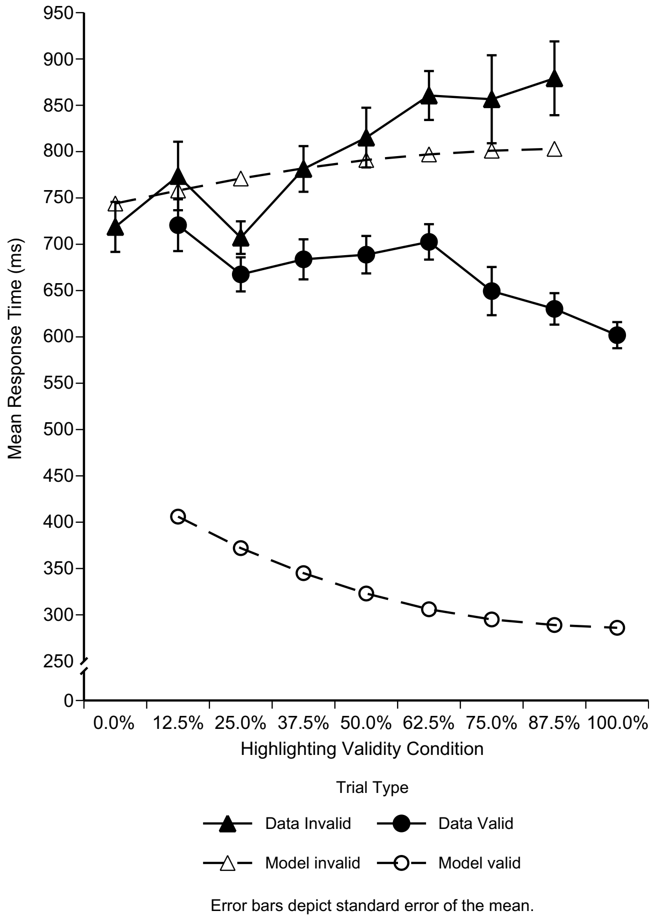

Fig. 4. Predicted RTs of Fisher et al.'s (1989) algebraic model and RTs from human data.

The algebraic 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 algebraic model to predict such fast times for valid trials, and it may be that subjects are not as fast on valid trials in low validity conditions because they are not so likely to search the highlighted item first.

Humans are probably not so likely to search the highlighted item first in Fisher and Tan's (1989) paradigm because they do not possess perfect knowledge of the environment's highlighting validity, as Fisher et al.'s (1989)

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

model does. In Fisher and Tan's paradigm, humans can only estimate highlighting validity by sampling it. Fisher et al.'s model does not consider any means of sampling probabilistic aspects of the environment. To better understand the nature of human cognition situated in tasks such as visual search of highlighted displays, a cognitive model must be built that incorporates some degree of probability sampling and learning. Probability sampling – testing the environment so that one may learn about it – could occur by occasionally acting contradictory to the optimal strategy – always attending the highlighted item first when validity is greater than 20% and never attending the highlighted item first when validity is less than 20%. When sampling probable aspects of an environment a human would occasionally attend first to an unhighlighted item even though the normative strategy dictates that the highlighted item should always be attended first whenever highlighting validity exceeds 20%.

What contributes to the adaptability of human visual search behavior in an environment where cues have some probability of being helpful or hindering? Firstly, there are probability sampling behaviors such as the behavior just discussed. Secondly, given a fairly simple paradigm like the digit search in a 5 × 1 array used by Fisher and Tan (1989), 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. Fisher et al.'s (1989) algebraic model does not consider this possibility. 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. It can be programmed to instantiate behaviors that sample probabilistic aspects of its environment and it can modify its behaviors accordingly. 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 just discussed.

2.4. 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"). "Attend-red" made ACT-R move visual attention to the red item. "Avoid-red" made it seek an unattended black item. If the currently attended item was a target, ACT-R output a press of the appropriate key.

If the item seen as a result of an attention shift was a distractor and the highlighted item had been attended, 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. This marking/not marking of "highlighted-distractor" was the sole and critical difference between the two models. If the red item was still unattended, then ACT-R 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).

At the start of each highlighted trial, the ACT-R models, visual-location buffer defaulted to a chunk describing the location of the red item. This made the location of the red item immediately available to the model and was meant to mimic the pop-out effect of visual salience. In the case of control trials, ACT-R had no red item to which to default the visual-location buffer, so it simply fixated unattended items until it found the target. Fig. 5 graphically depicts the algorithm used by both ACT-R models. The models' strategies for all cases was necessarily very simple because of the constraint imposed by rapid human participant response, approximately 700 ms on control trials. The 150 ms required for motor movement, the 185 ms required to decide to move visual attention, to move visual attention, and to encode the new object simply left no time for complicated cognitive strategies.

Production priors represent a method for ACT-R modelers to specify biases toward and away from certain behaviors that humans may bring to the task of interest (Anderson et al., 2004). Production priors for both ACT-R models 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

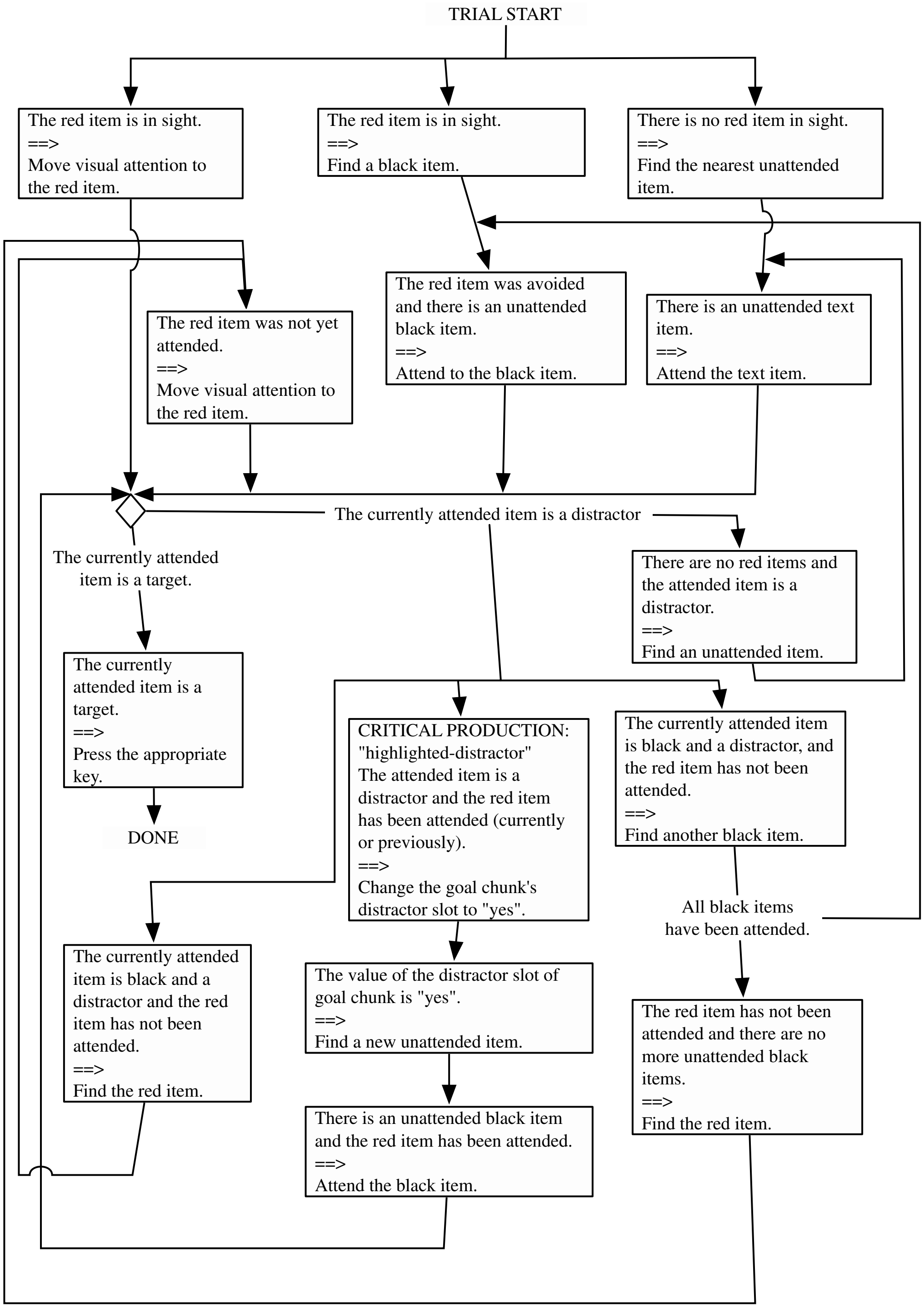
TRIAL START
The red item is in sight. ==> Move visual attention to the red item.
The red item is in sight. ==> Find a black item.
There is no red item in sight. ==> Find the nearest unattended item.
The red item was not yet attended. ==> Move visual attention to the red item.
The red item was avoided and there is an unattended black item. ==> Attend to the black item.
There is an unattended text item. ==> Attend the text item.
The currently attended item is a distractor
The currently attended item is a target.
There are no red items and the attended item is a distractor. ==> Find an unattended item.
The currently attended item is a target. ==> Press the appropriate key.
DONE
CRITICAL PRODUCTION: "highlighted-distractor" The attended item is a distractor and the red item has been attended (currently or previously). ==> Change the goal chunk's distractor slot to "yes".
The currently attended item is black and a distractor, and the red item has not been attended. ==> Find another black item.
All black items have been attended.
The currently attended item is black and a distractor and the red item has not been attended. ==> Find the red item.
The value of the distractor slot of goal chunk is "yes". ==> Find a new unattended item.
The red item has not been attended and there are no more unattended black items. ==> Find the red item.
There is an unattended black item and the red item has been attended. ==> Attend the black item.

Fig. 5. Algorithm employed by ACT-R models.

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

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 originally 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. 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 set 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.

One consequence of this set of priors is that the model is 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 Eq. (5), where P_i (Eq. (6)) 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 (Eq. (7)) is an estimate of the cost (typically measured in seconds) to achieve that goal (Anderson et al., 2004). "Successes" is the total number of times a goal has been achieved using production i, and likewise "Failures" is the total number of times a goal has failed to be achieved when production i has been applied. "Efforts" is the accumulated time (seconds) over all the successful and failed applications of production rule i. Successes, Failures, and Efforts can all be set to simulate prior experience. Thus "attend-red"'s utility started at $0.5 \times 20 - 0.05 = 9.95$ versus 15.90 for "avoid-red".

$$U_i = PiG - C_i \tag{5}$$

$$P_i = \frac{\text{Successes}}{\text{Successes} + \text{Failures}} \tag{6}$$

$$C_i = \frac{\text{Efforts}}{\text{Successes} + \text{Failures}} \tag{7}$$

ACT-R uses the utility of a production to decide which production to fire out of a set of competing productions (the conflict set) that match the conditions of the internal state of the model and the state of the outside world. The probability that production i will be selected on the basis of its utility, U, out of conflict set j is calculated with Eq. (8), where t controls the noise in the utilities. Summation is over all competing productions.

$$P_i = \frac{e^{U_i/t}}{\sum_j^n e^{U_j/t}} \tag{8}$$

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 to distractors are failures and only the attention shift to the target is a success. Learning only by time elapsed might not be enough feedback to change behavior in the macro model. If so, the model may need to register a failure every time it shifts attention to a distractor, as it does in the micro model.

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

Fig. 6 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 1000 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, $r^2 = 0.747$,

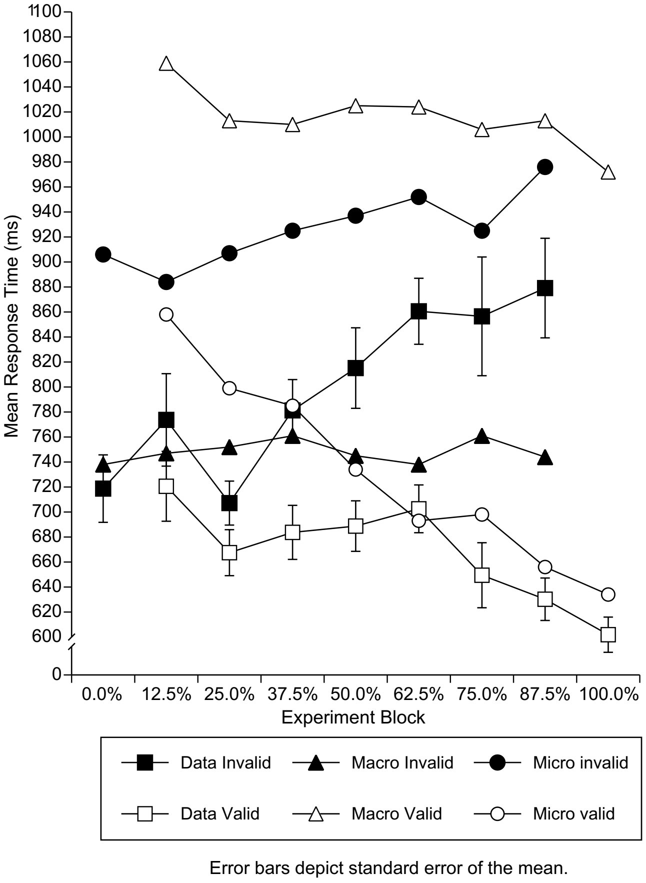

Fig. 6. Human data versus ACT-R models.

$p < 0.001$, mean deviation = 54 ms (Fig. 6). As highlighting validity increased, valid trials became faster and invalid trials became slower, as in the human data. It should be noted that the two models used the same production priors and same expected gain noise parameter, thus possibly not resulting in the best possible fit to data for each model. However, that fact only serves to illustrate our point: with only one change creating a local learning mechanism, the model's learning behavior radically changes in such a way as to enable it to go from being totally unable to mimic the human data to being able to at least duplicate the basic qualitative aspects, even if there are still problems in some conditions. The reasons for these fit problems are speculated upon in the general discussion section.

The one major flaw in the micro-level 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 specifically, 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 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 (Fig. 7), $r_2 = 0.911$, $p = 0.004$. 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 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 increases because it learns to attend to the highlighting more. However as validity increases, so does mean RT for invalid trials because the model more often attends to highlighting, including the occasions when highlighting is misleading.

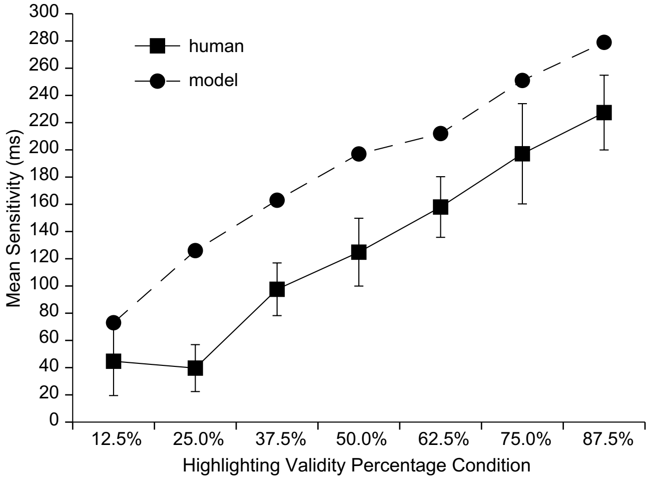

Fig. 7. Mean sensitivity for micro-level model and human data, across all experiment blocks. Error bars depict standard error of the mean.

3. General discussion

Fisher et al.'s (1989) algebraic 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 overestimates the degree to which people do check that group first. The Fisher et al. model failed because it did not consider human procedural learning situated in a probabilistic environment, particularly one so small.

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? People may be actively testing their environment so that they may learn about it. The algebraic model fails to predict human performance in a small search space because it does not take into account the fact that people are sensitive to the probabilistic structure of their environment.

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 the Fisher and Tan (1989) task those probabilities matter for costs and rewards because a shift of attention to the wrong item leads to 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., 2006), and that may be why the model fits poorly in the low highlighting validity percentage conditions. Gray et al. claim that a reinforcement production utility learning mechanism, "approximates what would be expected if human cognition calculated costs as if milliseconds mattered". In a world where individuals trials typically complete in three-quarters of a second, milliseconds clearly do matter.

However, despite the aforementioned shortcomings, the micro-level ACT-R model does exhibit sensitivity to the probabilistic aspects of its environment to a similar degree as the human subjects. ACT-R's reinforcement mechanism for learning production utilities is a promising and very recent development. At this point it would be speculative to make any claims about its performance in a task like the one reported in this study, and such investigation would represent a good next step for this project.

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 must 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. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, 113*(3), 461–482.

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.

Wolfe, J. M., Kluender, K. R., Levi, D. M., Bartoshuk, L. M., Herz, R. S., Klazky, R. L., et al. (2006). *Sensation and perception*. Sunderland, MA, USA: Sinauer Associates, Inc.