Modeling Icon Search in ACT-R/PM

Michael D. Fleetwood and Michael D. Bryne
{fleet, byrne}@rice.edu)
Rice University, Department of Psychology
6100 Main Street MS-25, Houston, TX 77005 USA

Abstract

As the use of graphical user interfaces expands into new areas, icons are becoming an increasingly important aspect of GUIs. Oddly, little research has been done into the the costs and benefits associated with using icons. One aspect of icons, icon borders, has been proposed as a means of adding information to icons. An experiment was conducted in which the potential cost in response time of using simple icon borders was investigated. Two models were then constructed in ACT-R/PM to carry out the same icon search task as in the experiment. The results of the modeling effort indicated an area where the design of the experiment could be improved. A second, "improved" experiment was carried out, the results of which suggest areas for further improvement in the ACT-R/PM models.

Introduction

Graphical icons are a standard feature of most graphical user interfaces (GUIs), representing a range of commands and objects. Usage of icons is, if anything, on the rise as small, low-resolution displays (e.g. mobile phones, PDAs), which make display of large amounts of text difficult, proliferate. Unfortunately, there is strikingly little empirical and theoretical work on how users interact with icons. A clearer understanding of the ways in which users search iconic displays could be of great value to designers of such interfaces, possibly allowing the display of more information in less space. The studies in this paper are aimed at examining the visual search processes employed by users as they search displays of labeled icons, a task we call *icon search*.

One aim of this particular line of research is to better understand how iconic displays can convey more information without penalizing the user by increasing search time. This was done by exploring the impact of adding borders to icons. A second goal is to explore the use of computational modeling in this critical task domain. We would like, ultimately, to be able to predict search times of iconic displays on an *a priori* basis through the use of computational models. This is an iterative, symbiotic process (Gray & Altmann, 1999). That is, we use models to gain greater insights into an applied HCI problem, while at the same time using the applied problems to identify areas where the models and modeling architecture can be improved. In that spirit, the first experiment was aimed at understanding an applied issue, the impact of simple icon borders. We then constructed ACT-R/PM models of the experiment, which in turn drove us to revise the experiment. Results from that experiment indicate shortcomings in the model and suggest places where our view of the strategies employed by users require revision.

1. Experiment 1

One aspect of icons that is ripe for improvement is the use of icon borders (Houde & Salomon, 1993). Currently icon borders are typically nothing more than a simple rectangle surrounding the graphical icon. Used in this manner, icon borders do little to convey additional information to the user. It may be that they only serve to add visual clutter to the display, and to provide another common element of targets and distractors, thereby potentially slowing down visual search (Mohr, 1984). It would be valuable to the graphical user interface (GUI) design community to have some understanding of the costs, if any, associated with using icon borders in this manner. It would also be valuable to gain some understanding of the relative efficiency or inefficiency of icon search in general. The following experiment was designed with these research goals in mind.

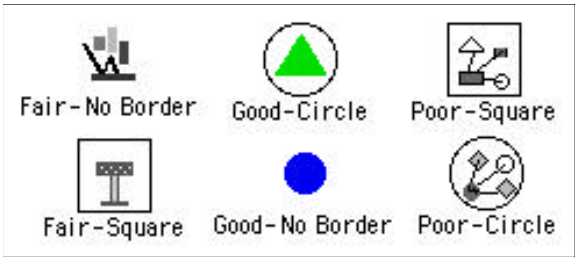

Figure 1. Some examples of the icons used in the experiments.

1.1 Method

1.1.1 Users

The users in the experiment were 20 undergraduate students at Rice University who were participating in order to meet a requirement for a psychology course.

1.1.2 Design

Three independent variables were manipulated, all of which were within-subjects factors. The first of these factors, set size, had four levels, 6, 12, 18, or 24 icons. A second within-subjects factor, target border, had three levels. The target icon to be searched for could be presented without a border (no-border condition), with a circle as a border (circle), or with a box as a border (square).

The final within-subjects factor, icon quality, had three levels. Icons were designed that varied in their level of

distinctiveness. On one end of the spectrum were icons of "good" quality. These icons were designed to be easily distinguishable from other icons based on the basic visual ("pop-out") features of color and shape (specifically curvature). Icons in the good quality set were one of six colors (red, blue, green, yellow, brown, or black) and one of two shapes (circle or triangle). On the other end of the quality spectrum were icons that were not easily distinguishable (referred to as "poor" quality icons). They were designed to be distinguishable in a set of two icons, but quite indistinguishable in a large distractor set. These poor quality icons were all of the same basic shape and did not include color (other than white, black and shades of gray). The "fair" quality icons were designed to be representative of the area in between these two ends of the spectrum. They were generally of a distinct shape, although more complex than the simple circles and triangles in the good quality icons, and none of them contained any color outside of the spectrum of gray scale colors. Refer to Figure 1 for examples of borders and quality levels. Each block in the experiment consisted of 36 trials (4 set sizes x 3 borders x 3 qualities).

The dependent variable being measured was the response time of the users—specifically, the time from when they clicked on a "Ready" button to indicate that they were finished examining the target icon to when they clicked on the target icon among the set of distractor icons. When the "Ready" button was clicked, the button and the target icon were removed and the display containing the distractor icons was presented.

One potential independent variable that was held constant in this experiment was the number of icons matching the target in the search display. On each trial one-third of the icons in the search display had the same pictorial icon and matching border. Thus, ultimately the user was forced to differentiate among the icons by the filename.

The location of the target icon was randomly selected for each trial. Also randomly selected were the file names for the icons. The distractor file names and the target file names were randomly selected without replacement from a list of 750 names until the list was exhausted. At which time, the list was recycled.

Each user completed four blocks of trials in addition to the practice block for a total of 180 trials.

1.2 Results

In Figure 2, mean response times are presented as a function of set size and icon quality. Here, it is evident that as icon quality decreases (good to fair to poor), response times increase. This is confirmed by a significant main effect of quality, $F(2, 38) = 52.14$, $p < 0.001$. Also, not only are the three qualities significantly different, but the slopes of the lines appear to be different, as confirmed by a reliable quality by set size interaction, $F(6, 114) = 5.20$, $p < 0.01$.

Relevant to the hypothesis concerning borders, there was no effect of icon borders on search time, $F(2, 38) = 1.66$, $p = 0.20$, nor did borders interact with any other factors. This suggests that users can indeed ignore the icon borders, and this may be an effective method for conveying additional information without increasing search costs.

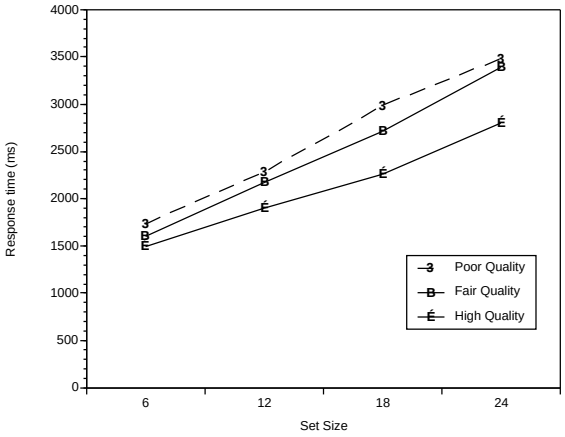

Figure 2. Mean response times by set size and icon quality, illustrating a main effect of icon quality.

2. Modeling the Experiment

We used ACT-R/PM (Byrne & Anderson, 1998) to model the experiment. Because the cognitive demands of the icon search task are minimal, modeling the perceptual-motor processes (e.g., shifting visual attention, pointing and clicking) with some fidelity is critical. The ACT-R/PM architecture combines ACT-R's theory of cognition (Anderson & Lebiére, 1998) with modal theories of visual attention (Anderson, Matessa, & Lebiére, 1997) and motor movement (Kieras & Meyer, 1997). ACT-R/PM explicitly specifies timing information for all three processes as well as parallelism between them.

2.1 The Task

The task performed by users seems simple: remember a target icon and filename, find it on a second display, and move the mouse to click on it. The first phase of this task, remembering the target icon and filename, is relatively straightforward to model in ACT-R/PM. The model attends the icon and selects a random element of the icon (e.g. "gray rectangle") to guide later search, which is noted in the goal chunk. The filename is also noted by storing it in the goal chunk. Searching for the target among the distractors in the second phase of each trial is more complex, and more than one strategy was implemented as described in the following section. Once the target icon was located, the model moved the mouse to the target and clicked on it.

2.2 The Models

Because our goal was to explore the space of strategies that users might employ, we constructed two models of the icon search task representing slightly different strategies. Both of the two strategies will be considered in some detail. Both of the models follow the same basic control structure but they differ slightly in strategy used to locate the target icon amongst distractors.

2.2.1 The "Double-Shift" (DS) Model

The double-shift model is so named because it requires two shifts of attention to examine each candidate icon. First, the model directs its visual attention to an icon that has at least one descriptive characteristic in common with the target icon (a red circle for example). This is done by finding a visual-location that contains the specific characteristic that was stored in the goal to "remember" the target icon. Then the model shifts attention to the filename directly below. This filename is compared with the target filename. If the filenames match, then visual attention is shifted back to the target icon so that it can be clicked on. If they do not match, then the process begins again by finding another icon on the distractor screen with the appropriate matching descriptive characteristic.

The quantitative predictions of this model are easy to compute in cases where there is no feature overlap between the target icon and the distractors, because the time parameters for all the operations are known and this model never re-visits icons on the display. For each icon examined, the model requires an additional 420 ms: one production to initiate the first shift to the icon (50 ms), time for the first shift (135 ms), one production to initiate the second shift to the filename (50 ms), time for the second shift (135 ms) and one production to do the comparison of the attended filename and the remembered filename (50 ms). Because there is no feature overlap in the "good" quality icons, and each set size adds two icons that match the target to the display (meaning on average one of them will be visited), the model clearly predicts a 420 ms slope for the RT by set size function for "good" icons. For other icon qualities, the slope depends on the degree of feature overlap between the target and the distractors. That is, if the target icon contained a gray square and the model selected that feature to guide search, then all icons on the display containing gray squares are candidates. The number of such icons will vary from trial to trial depending on the features in the target icon and the composition of the distractors, so Monte Carlo simulations are required to produce RT predictions.

2.2.2 The "Text-Look" (TL) Model

The text-look model is so named because attention is focused directly on the filename below the icon, and the actual icon is never actually attended. As in the double-shift model, an icon with a matching feature is located, but rather than shifting visual attention to the icon, it is shifted directly to the filename below the icon. This process is meant to simulate the process of preattentive search and the use of parafoveal vision by subjects. It is assumed in this model that under conditions where the target icon shares few features with the distractor icons—i.e. they are somewhat unique—then users do not need to examine the icon in detail, rather, they just look directly at the filename. The model does shift attention to the icon eventually, in order to move the mouse and click on it, but this attention shift only occurs after the target filename (and thus the target icon) has been identified.

Quantitative predictions from this model are not so easy to compute because this model may re-examine icons. This is because the icons themselves are never actually attended, and ACT-R/PM only "remembers" locations to which it has shifted attention. Because this revisitation is probabilistic, analytic predictions are difficult to derive and again Monte Carlo simulations are required.

2.3 Comparison of the Models

The two models represent slightly different strategies for the visual search. The DS model makes two shifts of attention per each additional icon examined, while the TL model makes only one, suggesting the TL model may be more efficient. However, the DS model does not revisit previously-seen items, which could make the DS model more efficient. We had no *a priori* predictions about which model would actually be faster.

The models have some key similarities as well. The production which selects the next icon to be examined selects randomly from all the candidates that match the remembered feature (e.g. "gray circle"). That is, there is no right-to-left or top-to-bottom pattern employed by the models. Because the location of the target was random, incorporating such a strategy would have made little difference in the ultimate predictions of the model in this experiment. Furthermore, the models employ the same strategy for all set sizes and icon qualities. These properties will be discussed further later in the paper.

Finally, both models depend on the representation of the icons themselves. Each icon is "seen" by ACT-R/PM's Vision Module as a list of features. For the "good" quality icons, this list is a singleton; that is, each icon is represented by a single feature (e.g. "red circle"). In contrast, more complex icons will have a number of features and colors associated with them, gray triangles, white circles, etc. What makes these more complex icons "poor" icons in the experiment is not the number of features the icon has per se, but rather the number of features the icon shares with other icons in the distractor set. For example, many of the icons in the poor quality set have gray triangles and white circles. As a result, the model will often examine icons that do not match the target icon exactly, but rather only share one particular feature with the target icon. It is this overlap of features, or "similarity," that makes such icons poor icons in context. In contrast, the good quality icons have no feature overlap with other good quality icons, and thus, only icons exactly matching the target icon are examined by the model. The exact nature and number of the features used to represent each icon in the "fair" and "poor" conditions are free parameters in the models; however, we used the same feature sets for both the DS and TL models.

2.4 Simulation Results

The model data was gathered by running the models for 80 blocks of trials, which approximates the number of real blocks of subject data involved.

It is clear from Figure 3 that the performance of the model is quite similar to that of real subjects. The model and the data show some divergence at the largest set size. This is

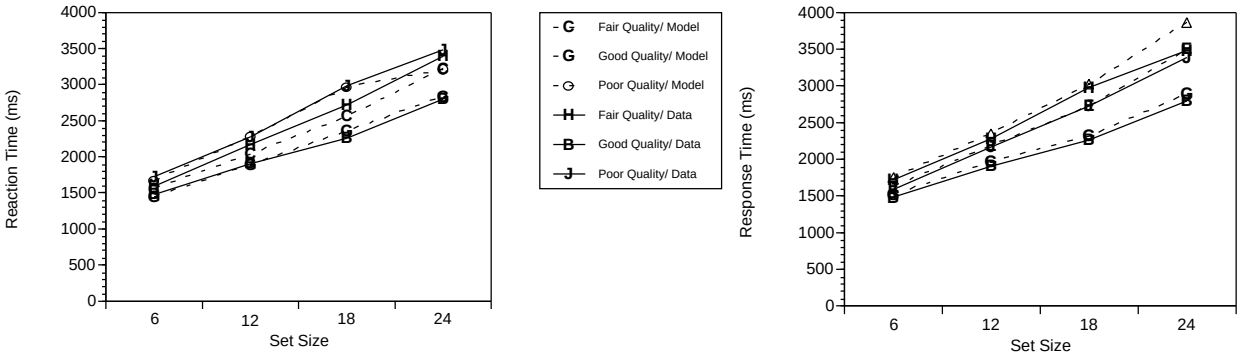

Figure 3. Mean response times by set size and icon quality for the TL model (left) and the DS model (right).

somewhat to be expected due to the greater amount of variability in reaction times at large set sizes. Most importantly, the model has retained each of the pronounced effects that were seen in the data—those of set size and iconquality. A final quantitative comparison of the two sets of data was obtained by examining the correlation and the percent mean deviation (root-mean-square) between the two sets of mean reaction times.

The R^2 between the data matrices for both the text-look and double-shift models is 0.98. The high correlation between the models and the data suggest that the models both do an excellent job of accounting for the major trends in the data. This provides some measure of validation for the models and the timing parameters incorporated into ACT-R/PM. However, these fits also depended on hand-tweaking of the basic visual features used to represent the icons, which will be discussed in the next section.

The root mean square error (RMSE) and percent average absolute error between the model and the data for the two charts shown previously are presented in Table 1. Note that the percent average absolute error in each of the above matrices of data remains in the remarkably low range of three percent—again indicating that the models were quite accurate in their performance in reference to the real subject data.

Table 1. Quantitative analysis of model data as compared to experiment data—root mean squared error and percent average absolute error.

	Double-Shift	Text-Look
RMSE	125.41ms	112.28 ms
Percent average absolute error	2.95%	3.19%

2.5 Discussion of Models

Based on the standard fit metrics, the fit of the models to the data is excellent for both models. The difference between the DS and TL models was very small, suggesting the increased number of attention shifts generated by the DS model were, in effect, cancelled out by the revisitation of the TL model. Based on the fits, there is no strong argument for preferring one model over the other, suggesting that strategy variation among users may not play a large role in determining search times. This was highly encouraging. However, there are two caveats. The first is that the feature lists used to represent each icon were generated *post hoc* to provide good fits of model to data. Ideally, we would like to have a principled way of constructing the feature lists on an *a priori* basis to make the models predictive rather than simply explanatory. This is an involved subject for future research.

The second caveat is that the DS model makes an analytic point prediction about the slope of the RT vs. set size function of 420 ms. What was distressing about this is that the observed slope *as well as the slope generated by the model* were approximately 460 ms. This was a surprise for us, and caused us to carefully re-examine the behavior of the model because it did not match our own analytic prediction. This led us to the discovery of an interesting point in the programming of the experiments. When the experiments were designed and programmed, we originally expected icon borders to play some significant role in the search strategies of users. One factor that we wanted to keep consistent in our design was the number of icons identical to the target icon except with different filenames (one-third of the icons in the set). However, because we assumed that icons with different borders would be seen as different than the target icon, we allowed icons matching the target icon except for the border to be included in the distractor set. As this experiment (and others we have conducted) have shown, borders such as those used in the experiment do not play a role in icon search. Thus, icons with different borders but with the same base pictorial icon may be seen as functionally the same icon. Because the distractor icons were chosen randomly, the experiment software occasionally (and randomly) allowed one or more additional (beyond the predetermined one-third) icons matching the target icon but with a different border to be present in the distractor set. The inclusion of this additional icon drove up reaction times for both users and the model, and was a

cause of the steeper than predicted slope.

While this was certainly a flaw in the design of the experiment, it speaks to a strength of the approach employed. The ACT-R/PM models interact with the same experimental software as human subjects, and thus are sensitive to the same factors. It was therefore encouraging to see that the model was affected by this flaw as well, and we certainly would not have noticed this flaw had it not been for the (mis)behavior of the DS model. As a result, in order to examine the model's predictions, and directly compare its results with those of users, we needed to remove this element of randomness from our design, and a follow-up experiment was conducted which did just that.

3. Experiment 2

By removing the aforementioned source of unpredictability in our program and, hence, a source of randomness or unexplained variance in our data, Experiment 2 was designed to get a "cleaner" picture of the data to allow for comparisons with a computational model.

3.1 Method

The design, procedures and apparatus of Experiment 2 were nearly identical to those of Experiment 1. The only factor that changed was the number of distractor icons in the distractor set that matched the target icon. In Experiment 2, distractor icons with the same base pictorial icon, but with a different border, were allowed to be randomly selected for the distractor set. In this experiment, icons with the same base pictorial icon as the target icon were excluded from the distractor set. The users in the experiment were 20 undergraduate students at Rice University who were participating in order to meet a requirement for a psychology course.

3.2 Results

In Figure 4, mean response times are presented as a function of set size and icon quality. Here again, as in Experiment 1, it is evident that as icon quality decreases (good to fair to poor), response times increase. This is confirmed by a significant main effect of quality, $F(2, 38) = 58.71$, $p < 0.001$. Also, not only are the three qualities significantly different, but the slopes of the lines appear to be different, as confirmed by a reliable quality by set size interaction, $F(6, 114) = 6.89$, $p < 0.01$. Additionally, as in the first experiment, there was no effect of icon borders on search time, $F(2, 38) = 1.10$, $p = 0.34$, nor did borders interact with any other factors.

3.3 Discussion of Experiment 2

Clearly, this experiment was able to produce the same general results from Experiment 1—effects of icon quality and set size and a lack of an effect of icon borders. Such replication allows us to draw our conclusions with even greater confidence.

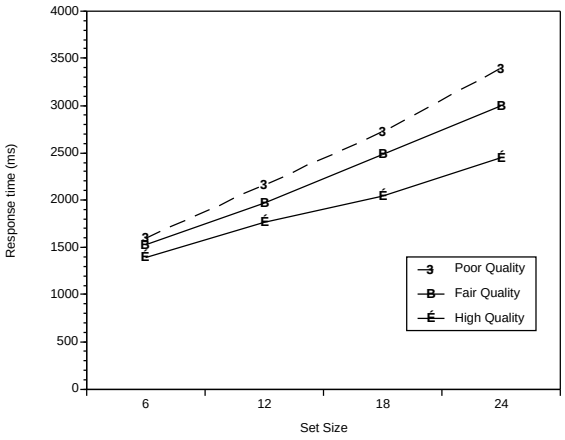

Figure 4. Mean response times by set size and icon of quality, illustrating a main effect of icon quality.

The particular aspect of this experiment that we were most interested in was not, however, these broader effects. We were interested in closely examining the effect of set size on the high quality icons. Specifically, we wanted to study the slope of the reaction time by set-size function for the good quality icons and determine if it approximated 420 ms, the value predicted by the DS model.

At first glance, the slope of the line appears to be shallower than 420 ms. The specific data points that make up the line are presented in Table 2 below. The average slope for the four data points was calculated as 355.08, quite a bit lower than 420 ms.

Table 2. Mean Reaction Times for Good Quality Icons across set size.

Set Size	6	12	18	24
RT (ms)	1385	1765	2042	2451

We computed the slope for each subject and generated the 95% confidence interval for the slope, which covered 311 ms/icon to 392 ms/icon. Thus, the model's analytic prediction of 420 ms is clearly an overestimation of the true slope. This is a significant misfit of the data by the model, and provides the impetus for a revision of the model.

Section 4: General Discussion

Our model, in fact, is too slow. Real subjects can find the icon faster than our model can under optimal conditions of zero feature overlap. The solution then is to change the model, but that is certainly easier said than done. There are numerous ways that the strategy of real subjects may differ from the strategy implemented by the model. It would make sense then to investigate what the real strategies of users are in the icon search task. In order to do this, one would have to actually study users' eye movements while engaged in the task, accomplished through the use of an eye tracker.

Our current research is therefore concentrated on eye tracking the icon search task. Initial data have been collected, and although they have yet to be quantitatively analyzed, two aspects of users' search strategies are apparent from a purely qualitative look at the data. For one, it seems that users employ different search strategies based on the quality of the icons. For example, one strategy employed in the good quality condition is a "global grouping effect." In this condition, it appears that users are able to identify clusters of icons preattentively. Having identified a group of icons that match the target icon, they begin their search in this group. In this manner, they do not follow a simple left to right or random strategy of examining icons that match the target icon. In contrast, in the poor quality condition, there is no apparent directed strategy to any group of icons or area of the screen. Some of the users' quicker response times, relative to the model response times, may be due to such a directed and superior strategy.

Another pattern that is suggested by the data is potentially that a strategy approximating the text-look strategy is more frequently employed by users than the double-shift strategy. It appears that users rarely look directly at the icon before looking at the filename below it. However, even the text-look strategy falls short of the behavior of users. Often, they do not have to even look at the filename in order to reject it as a possible candidate for the filename they are searching for. For example, users seem to be able to reject certain filenames based on the length of the word. If they are looking for a short filename they can reject a long filename without actually reading it and vice versa. In order to extract such information from the display without foveating the source of the information, users must be able to extract information parafoveally.

We would like to be able to incorporate these two findings into an improved version of the icon search model. ACT-R/PM currently assumes a direct correspondence between unobservable attention shifts and observable eye movements; that is, people fixate the target of attention. Such an assumption holds in some cases, but it is agreed upon in the research community that it does not hold in general (Henderson, 1992; Rayner, 1995). The experiments modeled here may provide an example of one such case where this does not hold. Therefore, in order to model the experiment accurately in ACT-R/PM, some of the underlying assumptions of the Vision Module need to be improved upon. Fortunately, there already exist computational models that serve as a bridge between observable eye movements and the unobservable cognitive processes and shifts of attention that produce them, such as EMMA (Eye Movements and Movements of Attention) developed by Salvucci (2000).

Such a recommendation for improving ACT-R/PM illustrates the symbiotic, iterative nature of empirical experimentation and computational modeling that was alluded to previously. We initially turned to modeling in order to get a better conception of an applied task and some experiments dealing with that task, icon search. Indeed, the modeling effort proved valuable in this respect, even pin-pointing where the conceptual design of our study needed to be improved. However, the implementation of these improvements in our experimental design has brought to our attention an aspect of the Vision Module in ACT-R/PM that might be improved. Our next step is to turn to creating a new model of the icon search task based on the information in the eye-tracking study.

References

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

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interacton*, 12, 439-462.

Byrne, M. D. & Anderson, J. R. (1998). Perception and Action. In J. R. Anderson & C. Lebiere (Eds.) *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Gray, W. D., & Altman, E. M. (1999) Cognitive modeling and human-computer interaction. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors.* New York: Taylor & Francis, Ltd.

Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In K. Rayner (Ed.), *Eye Movements and Visual cognition: Scene Perception and Reading*. New Your: Springer-Verlag.

Houde, S. and Salomon, G. (1993). Working towards rich and flexible file representations. *Proceedings of the CHI '94 conference companion on Human factors in computing systems*.

Kieras, D. E., & Meyer, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, 104, 3-65.

Mohr, W. (1984). Visuelle Wahrnehmung und Zeichenfunktion, (Roderer, Regensburg); cited in Scott, 1993.

Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. In J. M. Findlay, R. Walker, & R. W. Kentridge (Eds.) *Eye Movement Research: Mechanisms, Processes, and Applications.* New York: Elsevier Science Publishing.

Salvucci, D. D. (2000). *An integrated model of eye movements and visual encoding.* Submitted manuscript.

Scott, D. (1993). Visual search in modern human-computer interfaces, *Behaviour & Information Technology*, 12, 174-189.