Strategies for Guiding Interactive Search: An Empirical Investigation into the Consequences of Label Relevance for Assessment and Selection

Duncan P. Brumby

Andrew Howes

Cardiff University

RUNNING HEAD: STRATEGIES FOR INTERACTIVE SEARCH

Paper to appear in *Human-Computer Interaction*.

Corresponding Author's Contact Information:

Duncan Brumby

Department of Computer Science

Drexel University

3141 Chestnut Street

Philadelphia, PA 19104

Email: Brumby@cs.drexel.edu

Brief Authors' Biographies:

Duncan Brumby is a cognitive scientist with an interest in human-computer interaction, eye-tracking, and computational cognitive modeling; he is an Associate Research Scientist in the Department of Computer Science at Drexel University. **Andrew Howes** is a cognitive scientist with an interest in the constraints on cognitive adaptation; he is a Reader in the Manchester Business School, University of Manchester.

Abstract: 254

Total number of words: 16,500 (not including references, tables, and figures)

ABSTRACT

When searching a novel web page people often estimate the likelihood that labeled links on the page will lead to their goal. A rational analysis of this activity suggests that people should adjust their estimate of the likelihood that any one item will lead to the goal in a manner that is sensitive to the context provided by the likelihoods that other items on the page will lead to the goal. Two experiments were designed to provide evidence to discriminate between this account and others found in the literature (e.g., satisficing and assess-all accounts). The experiments systematically manipulated the relevance of the distractor items and the location of the target item on the page. The results showed that (1) a high value item was more likely to be selected when it was first encountered if the relevance of competing distractors was relatively low and (2) that more items were assessed prior to selection when the distractors were of greater semantic relevance to the goal. The location manipulation showed that if more distractors were assessed prior to the target item, then the relevance of the distractors had a greater influence on the decision as to whether to select the target immediately. These results suggest that decisions as to when to select an item from the page are sensitive to the context provided by the likelihoods of all of the items so far assessed, and not just to the most recent item. The findings are therefore inconsistent with both satisficing and assess-all accounts of interactive search.

CONTENTS

1. INTRODUCTION

How do people search a newly encountered web page for a link that is relevant to the achievement of their search goal? It is known that estimates of label relevance play a substantial role in determining link selection (Card, Pirolli, Van Der Wage, Morrision, Reeder, Schraedley, & Boshart, 2001; Chi, Rosien, Suppattanasiri, Williams, Royer, Chow, Robles, Dalal, Chen, & Cousins, 2003; Katz & Byrne, 2003; Miller & Remington, 2004). Theories of label relevance have been partly motivated by the hope that the usability of a web site might benefit from improving the information architecture of a web site (Blackmon, Kitajima, & Polson, 2005, 2003; Blackmon, Polson, Kitajima, & Lewis, 2002; Chi, Pirolli, Chen, & Pitkow, 2001; Chi, Pirolli, & Pitkow, 2000; Kaur, & Hornof, 2005). One neglected issue, however, concerns the fact that during a search the user must decide which items to assess. In other words, estimates of relevance must be embedded within a strategy for controlling search.

We refer to the activity of searching a novel web page for information that is relevant to the achievement of a particular search goal as interactive search (Payne, Richardson, & Howes, 1999). While there is little empirical evidence about how people control search during such tasks, a number of cognitive models of interactive search have been proposed (Brumby & Howes, 2004; Cox & Young, 2004; Howes, 1994; Howes, Payne, & Richardson, 2002; Lee & MacGregor, 1985; MacGregor, Lee, & Lam, 1986; Miller & Remington, 2004; Pirolli & Card, 1999; Pirolli & Fu, 2003; Rieman, Young, & Howes, 1996; Young, 1998). In each, quite different assumptions have been made about how people choose between assessment and selection. The studies presented in this paper were aimed at discriminating between these assumptions.

In some accounts it is assumed that people tend to consider all of the items on a page prior to making a selection, whereas in others that people make a selection immediately following an assessment of a highly relevant item. The former assumption has the advantage that it guarantees that the label with the highest likelihood will be selected. The latter is more like Simon's satisficing heuristic (Simon, 1955). It has the advantage that a good enough label may be found in less time. In other accounts it is assumed that people consider the costs and benefits of further assessment in the context of the information that they have so far gained from the current page (Cox & Young, 2004; Young, 1998). These normative models are inspired by Anderson's (1990) rational analysis. In Young's rational analysis of exploratory learning (a general class of tasks that includes interactive search), the assessment of a choice carries a time cost, but also carries the potential benefit that the information gained will reduce the risk of incurring the cost of an erroneous selection. Young's analysis suggests that people should neither assess every item in a menu, nor select an item immediately after a positive assessment, rather they should continue to assess until it is rational to make a selection. The point at which

selection is rational is determined by a set of factors that include the context provided by previous assessments. We review these models in more detail below.

The paper is organized as follows. We first discuss previous empirical studies of interactive search that have focused on regularities in how people search web pages and database menus. We then review the substantial theoretical literature concerning how search behavior might be controlled.

Two experiments were designed to provide evidence to discriminate between the competing accounts of interactive search behavior. With Experiment 1 we systematically manipulated the relevance of the distractor items in the choice set, and with Experiment 2 we manipulated the location of the target item within the set. The experiments demonstrate that the likelihood that a person selects an item depends not only on the most recent assessment of that item, but also on the quality of the entire set of assessments made so far. The contribution of this work furthers our understanding of how people examine labeled links on a web page during goal-orientated search.

1.1. How Do People Search Web Pages?

While our focus in this article is very much on the consequences of label relevance for the strategies by which people choose to assess and select web links, there is a large literature on the topic of how people search web pages and sites. For example, people use the web to fulfill a variety of everyday needs (Byrne, John, Wehrle, & Crow, 1999; Cockburn & McKenzie, 2001; Morrison, Pirolli, & Card, 2001; Sellen, Murphy, & Shaw, 2002). They use it by both navigating links and by using a search engine. They often fail to go directly to a site or page that satisfies their goal, and as a consequence, they use the backup button quite frequently (Catledge & Pitkow, 1995). The design of a site (e.g., its depth/breadth) can affect how quickly people satisfy goals (Katz & Byrne, 2003; Larson & Czerwinski, 1998; Parush & Yuviler-Gavish, 2003; Miller & Remington, 2004; Norman, 1991; Snowberry, Parkinson, & Sisson, 1983). The spatial layout of the page has consequences for ease of navigation (McCarthy, Sasse, & Riegelsberger, 2003), as does the color of the hyperlinks (Halverson & Hornof, 2004; Pearson & van Schaik, 2003). This literature provides an invaluable context against which the role of relevance in guiding interactive search should be set.

The relevance of links to a user's particular information goal is one issue that has received considerable empirical and theoretical attention (Blackmon et al., 2005, 2003, 2002; Katz & Byrne, 2003; Card et al., 2001; Chi et al., 2003, 2001, 2000; Kaur, & Hornof, 2005; Miller & Remington, 2004; Pierce, Parkinson, & Sisson, 1992; Pirolli & Fu, 2003). Unsurprisingly, users tend to select items from a web page that are relevant to their goal (Katz & Byrne, 2003; Card et al., 2001; Miller & Remington, 2004). Card et al. observed participants while engaged in goal-directed search of the web for sites relevant to achievement of ecologically determined goals. A user trace was constructed based on eye-tracking data, application-level logs, and verbal protocols. They found that

participants were likely to select items from a web page (i.e., labeled links) that were of greater semantic relevance to their information goal. This scent following strategy was particularly evident from verbal protocols. It was also found that when a participant traversed a number of pages within a web site they tended to switch to a different web site when the relevance of the items in the site decreased below some typically experienced value. In a study by Katz and Byrne participants searched toy-web sites to locate links that were relevant to a given goal statement. In the experiment participants could opt to navigate the site by either using an inbuilt search feature or by browsing a menu structure. Katz and Byrne found that the decision between using the search feature or browsing a menu structure was influenced by the number of labeled links on the page and the semantic relevance of the links to the users search goal; when a page contained more labeled links of lower relevance to the search goal, participants were more likely to use the search feature.

Although users tend to select items that are relevant to their goal, one important issue to consider is the consequences of the ease with which they can successfully discriminate goal-relevant target items from the surrounding distractor items. Obviously, the first consequence is that more incorrect items may be selected. This fact can interact with the structure of the site. Miller and Remington (2004) conducted a study that manipulated label relevance and the information structure of a web site (breadth vs. depth). The depth-breadth trade-off considers the optimal page arrangement for the set of labeled links within a site. In a deep information structure there are many levels of pages in the structure but each of the individual pages contains few items. In contrast, a broad information structure has fewer levels of pages but each of the pages inevitably contains many more items. Miller and Remington found that participants were quicker at finding a target node with a deep information structure when the items on the route to the target node were clearly discriminable from the surrounding distractor items at each page on the path to the target. When the path to the target was not clearly discriminable, however, a broad information structure gave faster search times.

The sensitivity of web search behavior to label relevance has been further investigated with analytic techniques that predict search behavior on the basis of statistics derived from text corpora (Blackmon et al., 2005, 2003, 2002; Chi et al., 2003; 2001, 2000; Kaur & Hornof, 2005). The aim has been to provide estimates of relevance given an information goal and the content of a web page. The most common statistical approaches that have been used include Latent Semantic Analysis (LSA, Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997), and Point-wise Mutual Information using Information Retrieval (PMI-IR, Farahat, Pirolli, & Markova, 2004; Turney, 2001). Kaur and Hornof (2005) offers a comparison between the performance of these and other semantic systems in predicting the link that people would select given an information goal and a web page. Chi and colleagues (Chi et al., 2003; 2001, 2000) have developed usability tools that aim to predict the items on a web page that users are likely to select for a particular information goal. Blackmon and colleagues

(Blackmon et al., 2005, 2003, 2002) have also proposed techniques by which the usability of a web site may be improved by analyzing the relevance of the labeled links on a page.

While usability tools that aim to predict which items on a web page users are likely to select have assumed that the relevance of all of the links on a page are considered prior to selection, there is no evidence reported to support this assumption. In fact, very early work provided evidence that while people sometimes assess every possible link prior to selection, in many cases they often do not. We next review empirical studies that have examined peoples' strategies for controlling interactive search.

1.2. Strategies for Controlling Interactive Search

A study by MacGregor, Lee, and Lam (1986), that predates the invention of the web, observed a range of interactive search behaviors. In the experiment participants searched a database by selecting items from menu pages. The experiment manipulated the number of items that were presented on each menu and whether the participant could see all the menu items at the same time (simultaneous search) or only a single menu item at a time (sequential search). This latter sequential search condition allowed for participants search behavior to be inferred based on the number of items the participant chose to uncover prior to the selection of an item. MacGregor et al. observed three behaviors, which the authors labeled: *self-terminating*, *exhaustive*, and *redundant*. The self-terminating behavior consisted of a participant looking at and evaluating each item in turn until one was examined that was considered sufficiently relevant that it was selected immediately. The exhaustive behavior was evident when people first looked at and evaluated all of the menu items and then returned to and selected the one with the best evaluation. The redundant behavior consisted of repeatedly looking at and evaluating some subset of the items before making a selection. None of the participants in the study consistently exhibited a single search behavior and two-thirds of the participants showed all three. Furthermore, MacGregor et al. observed that participants' search strategy was contingent on the size of the choice set; as the number of items increased participants were more likely to self-terminate.

In a similar study, Pierce, Parkinson, and Sisson (1992) considered how the semantic relevance of the items in the menu affected search behavior. The experiment used a similar methodology as MacGregor et al. (1986), where participants searched single-page menus in which the semantic relevance of a target item was varied. Pierce et al. found that when the target item was less semantically relevant to the goal statement, the participants were more likely to exhibit an exhaustive or redundant search behavior and were less likely to accurately select the target item. When the target item was highly relevant to the search goal, however, the participants were more likely to self-terminate by selecting it without assessing any further items.

Studies of people learning computer application menus, rather than searching database menus, are also relevant to scoping the range of possible interactive search strategies. For instance, Franzke (1995, 1994) and Rieman (1994) found that participants demonstrated a label following strategy while learning a novel graphing package (called *Cricketgraph*), in that they tended to select items from the application menu with labels that had a high semantic overlap with the current goal. Rieman gained further understanding of participants' exploratory behavior by focusing on the search behavior leading up to the selection of an item. Analyses of verbal protocols and mouse movements suggested that prior to the selection of an item participants would often not assess all of the items in the available choice set and would repeatedly reassess an increasingly small subset of those items that were initially assessed. Participants also invested more time on each successive assessment of an item. Rieman, Young, and Howes (1996) later characterized this search strategy as an *iterative deepening of attention*, involving the progressive focusing on a set of potential items with greater effort placed in thinking about the meaning of an item's label on subsequent passes. This search behavior does not fit into the exhaustive, redundant, or self-terminating taxonomy proposed by MacGregor et al. (1986) for single-page menu search. Moreover, the search behaviors observed in the studies by Rieman (1994) and Franzke (1995, 1994) suggest that multiple assessment methods are deployed during interactive search.

The behaviors observed when people search menu pages in a database (MacGregor et al., 1986; Pierce et al., 1992) or learn a computer application menu (Franzke, 1995, 1994; Rieman, 1994) suggest different ways in which people may control interactive search. It is an open question, however, whether people adopt similar behaviors during web-based interactive search. First, the content of database menus and computer application menus are usually substantially different to the content of web pages. Second, these studies used invasive methodologies to infer participants search behavior. MacGregor et al.'s experiment used a sequential presentation methodology to determine the number of items that participants chose to assess prior to selection. Lohse and Johnson (1996) have demonstrated that this methodology can substantially alter information acquisition behavior. Franzke's and Rieman's studies of people learning application menus used a potentially invasive talk-aloud protocol methodology. These methodological issues are described further below.

Our aim in this paper is to expose the details of the eye-movement strategies, and by inference the assessment strategies that people use during interactive search. It is necessary to first review the substantial theoretical literature because various predictions concerning these strategies, some which go beyond the data, can be derived.

1.3. Computational Cognitive Models of Interactive Search

There have been a number of models of the cognitive processes that might be involved in controlling interactive search (Brumby & Howes, 2004; Cox & Young, 2004; Howes, 1994; Howes et al., 2002; Lee & MacGregor, 1985; MacGregor et al., 1986;

Miller & Remington, 2004; Pirolli & Card, 1999; Pirolli & Fu, 2003; Rieman et al., 1996; Young, 1998). In general, people are assumed to be sensitive to some form of estimate of the likelihood that a labeled item will lead to the goal; however, the models differ in terms of which items are considered and in the selection strategy adopted. One dimension on which the models differ concerns the assumptions that are made about how people choose whether to select an item or continue assessing items (i.e., what people are sensitive to when searching for information).

Information foraging theory (Pirolli & Card, 1999; Pirolli & Fu, 2003; Pirolli, 2005) has had a seminal contribution in building our understanding of how people search the web. The theory assumes that during web-based information gathering activities people are sensitive to the rate of information gain in relation to the cost of interaction. The theory assumes that the relation of navigation cues (information scent) to the user's information goal determines browsing actions.

A cognitive model, called SNIF-ACT (Pirolli & Fu, 2003), has been developed based on information foraging theory and implemented in a modified version of the ACT-R cognitive architecture (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004). SNIF-ACT uses a spreading activation mechanism to assess the relevance of items on a page. A key contribution of this work is the hypothesis that people will leave a site/page when the rate of gaining information falls below the average rate of gain. One of the assumptions that was made in the SNIF-ACT model, however, was that the likelihood of *every* item on a page is considered prior to selection. Therefore, while site-leaving decisions are sensitive to the economics of information gain, the choice of which items to assess is not. In other words, an *assess-all* decision strategy is used in SNIF-ACT to control within page. This simplification is potentially non-trivial if, as is suggested by MacGregor et al.'s (1986) menu search data, people sometimes choose to select an item without assessing any further items in the choice set (i.e., self-terminating search).

To account for their empirical findings MacGregor et al. (1986) described a model of single-page search in which the decision as to whether to select an item, assess a new item, or reassess an existing item was sensitive to the value of the most recently assessed item relative to a threshold. A number of behaviors were captured with this threshold strategy. First, if the likelihood that an item would lead to the goal clearly exceeded a selection threshold then there was a chance that the item would be selected immediately, without further evaluation. Second, if an item was only just above the threshold then it would be considered as a possible choice but not selected immediately; therefore, further items in the set would be evaluated. Finally, if after having examined all of the items, more than one just exceeded the threshold then the model re-examined this subset.

Miller and Remington (2004) proposed a cognitive engineering model, called MESA, which simulated navigation of a multi-page web site. Similar to MacGregor et al.'s (1986) threshold model, Miller and Remington also assumed that the relevance of items on a web page were assessed and that an item was selected if it exceeded a relevance

threshold. If when all the items on the page were assessed none had exceeded the threshold, then the threshold was lowered and the items on that page reevaluated relative to the new, lowered threshold. The model returned to the previous page in the site when none of the items on the page exceed the reduced threshold.

Howes, Payne, and Richardson (2002) reported a model in which the decision to leave a page (i.e., to select a backup button) was moderated, not only by the relevance of the items on the current page, but also by the relevance of items on the previously visited pages of the current site. Howes et al.'s model also used a simple selection threshold, but importantly, the threshold was dynamically determined by the distal search context. Consistent with Payne et al.'s (2000) empirical data, the model used an episodic memory of previous assessments to determine whether the utility of backing up was greater than the utility of an available forward move.

Rieman, Young, and Howes (1996) proposed a model called IDXL. The model captured Rieman's (1994) earlier observation that multiple assessment methods are deployed during interactive search. The model searched both an external menu and the internal space of possible evaluations and was sensitive to the costs and benefits of different methods of assessing items. The model evaluated menu items in turn, starting with a relatively low cost evaluation of the menu items, and moving to a more sophisticated, but higher cost, assessment procedure. A low cost assessment might be characterized by, "*Does the currently attended item contain a word that is also in the explicitly articulated goal description?*" Applying this assessment procedure sometimes identified items that provided exact label matches with the goal description, which resulted in the selection of an item. If none of the items provided an exact label match with the goal description, then the model would reassess a subset of the menu items with a higher cost, but more sophisticated assessment procedure, such as "*Is there a semantic link between an items label and the goal?*" The model exhibited behaviors consistent with observations of participant learning computer applications menus (e.g., Franzke, 1995, 1994; Rieman, 1994): exact label matches were selected sooner than labels that were synonyms of the goal description, and the model would reevaluate a subset of the available menu items with increasing attention on each successive pass. Moreover, IDXL went beyond MacGregor et al.'s (1986) model by embedding hypotheses about the details of the cognitive processing that is conducted during interactive search.

Young (1998), and more recently Cox and Young (2004), reported a rational analysis of exploratory learning in which assessments were used to reduce uncertainty concerning the likelihood that each item would lead to the goal. Following Rieman et al. (1996), multiple assessment methods of varying quality and cost could be applied to a given menu item to provide an independent assessment of the item's relevance, reflecting a subjective judgment of the likelihood that the selection of the item would lead to the goal. In Cox and Young's analysis it is assumed that at each step the choice between assessment methods is sensitive to a trade-off between the benefit of increased

confidence and the incurred cost of implementing the assessment. If the costs of two assessments were the same then the one that would be expected to most reduce uncertainty would be favored. The analysis suggests that it is rational to select the item from the menu with the greatest relevance estimate only when the expected reduction in uncertainty (i.e., the information gained) from further assessment is no longer worth the cost incurred. Cox and Young claim that a broad range of search behaviors are emergent from this single decision strategy.

A key assumption in Cox and Young's work (Cox & Young, 2004; Young, 1998) is that the current estimate of the likelihood that a given menu option will lead to the goal is dependent on all the other assessments that have been made. This assumption is represented mathematically by normalizing each estimate over the sum of all other estimates after each new assessment. As Young states, the assumption, "reflects real cross-relationships between the judgments about choices made by a person, and cannot be avoided … the reality is that people are often forced to make rapid and radical revisions of their estimates of the correctness of particular options as they work their way through [the options available]" (Young, 1998, pp. 474). Given that there is at least one item that will lead to the goal, it is rational to assume that a reduction in the probability that a particular item is the target will lead to an increased probability that some other item is the target. The model is particularly interesting because it makes a novel and empirically untested prediction. Hitherto an assumption that has been made in the literature is that estimates of the likelihood that an item will lead to the goal are independent of context; however, in Cox and Young's framework it is assumed that not only the relevance of an item affects the estimate of likelihood, but that the relevance of other examined items in the choice set also influences likelihood. In other words, the subjective value of selecting an item is sensitive to the context provided by the previously visited item in the choice set. We refer to models that adopt this assumption as context-sensitive accounts.

Brumby and Howes (2004; see Brumby, 2005, for more details) have presented an ACT-R (Anderson et al., 2004) model of interactive search that was influenced by Young's (1998) rational analysis. In the model the assessment of items involved the retrieval of item-specific chunks represented in declarative memory. The dependency between assessments was captured in the model by assuming that estimates of an item's likelihood were mediated, in part, by a pre-existing attentional focusing assumption in the ACT-R architecture. The ACT-R theory of declarative memory assumes that there is a fixed amount of source activation distributed between chunks in memory that are directly associated with the goal (or more precisely, the elements that are part of the ACT-R goal). This assumption was originally introduced as a constraint that operates over a spreading activation network (Anderson, 1983; Anderson & Pirolli, 1984) and is directly supported by an empirical study (Sohn, Anderson, Reder, & Goode, 2004), which demonstrated that manipulation of attentional focus affects the size of the classic fan effect (Anderson, 1974). Importantly, the fact that the amount of source activation available is fixed means

that if the activation of one goal-associated chunk is increased, then the total source activation received by the rest decreases. Chunks can also vary in how closely they are associated with the goal. Chunks with more, or stronger, associations with the current goal get more activation, and chunks with fewer, or weaker, associations with the goal get less activation. Attentional focusing thereby provides an architectural mechanism with which to model the idea that people will normalize (i.e., that they will rationally increase the likelihood that a choice is correct when another is rejected).

Brumby and Howes (2004) proposed that menu item assessment could be modelled with ACT-R attentional focusing. Menu items could be represented with chunks in declarative memory, called *item-chunks*, and assessment of an item could result in the addition or removal of associative links between the relevant item-chunk and the goal. When there are only a few links between the goal and item-chunks then the associations will be stronger. The more links that are added between the goal and the item-chunks then the more distributed and weaker each chunk's activation.

Initially, one item-chunk could be associated with the goal for each item in the menu. (If there are ten items in the menu then ten chunks would be associated.) Source activation would therefore be distributed equally to all items. Assessments of relevance based on examining item labels would lead to the addition or removal of associations. Associations would be added in the case of positive assessments and removed for negative assessments. Assessment of an item would lead to the removal of a link between the relevant item-chunk and goal if there was a failure to retrieve the chunk for that item.

Poor distractors are more likely to have associations removed early because of retrieval failures. Removal of a link between an item-chunk and the goal would mean that other items in the set are strengthened (because the fixed source-activation would be distributed less broadly). The relevance, and association, of any one item-chunk can thereby moderate the activation of other item-chunks. Lastly, Brumby and Howes, proposed that the decision to select an item was governed by a threshold that represented the benefit of selection. The item that eventually exceeded the threshold was selected. The use of attentional focusing ensured that the expected value of selecting an item was sensitive to the context provided by the previously visited items in the choice set.

While it made use of an available architectural mechanism, Brumby and Howes' (2004) model differed substantially from previous ACT-R models of interactive search. For instance, Pirolli and Fu's (2003) SNIF-ACT model of web search was developed in the ACT-R architecture, but focused on the economics of navigating sets of pages, rather than searching a single page.

1.4. Summary

It is generally accepted that estimates of label relevance play a substantial role in guiding navigation behavior on the web; however, there are important differences between the selection strategies that make use of these estimates. Some models have assumed what we shall refer to as an assess-all strategy, in which all of the items in a choice set are assessed prior to selection. Other models have assumed a simple-threshold strategy that is similar to a satisficing heuristic (Simon, 1955). Lastly, following Young (1998), there is a class of models that we refer to as context-sensitive. These latter accounts assume that people are sensitive to the cross-dependencies between the likelihoods that items in a choice set will lead to the goal. From this perspective, people choose to further assess items so long as the potential reduction in uncertainty as to which is the target item outweighs the cost of carrying out the assessment.

In the next section we report two empirical studies that aimed to evaluate these accounts. The experiments systematically manipulated the relevance of the distractor items and the location of the target item on the page. An eye-tracking methodology was used to determine the number of items in the set of options that participants tend to look at prior to selecting an item.

2. EXPERIMENT 1

The aim of Experiment 1 was to test the context-sensitive account by systematically manipulating the relevance of the distractor items in the choice set and measuring the consequences for which of the set of items were assessed. If decisions to select an item are sensitive to context, then participants should assess more items in menus that contain more highly relevant distractors. If participants use a simple-threshold strategy, however, then they should assess fewer items in menus that contain more highly relevant distractors (because on average one or other item will exceed the threshold earlier). Lastly, if participants tend to assess all of the items and then pick the best, then the relevance of the distractors should have no effect on the number of assessments made.

Specific predictions can also be made about reassessment. Cox and Young's (2004) model predicts revisits to items will be common, reflecting the application of assessment methods of varying cost and quality to items prior to selection. In this respect it is consistent with the related empirical (Franzke, 1995, 1994; Rieman, 1994) and theoretical (Rieman et al., 1996) work concerning how people learn to use a novel computer application interface through exploratory learning. In particular, it is predicted that participants will be more likely to revisit items that are more relevant to the goal description and to place greater effort into thinking about the meaning of an item's label on subsequent revisits. (Note: These predictions do not concern the number of eye movement fixations that will be made but the number of items that will be fixated and the duration of each visit to a particular item.)

Before describing the experiment, we first need to address the issue of how to determine which items a participant assesses. Previous studies of the strategies that

people use to search database menus used a process tracing methodology (e.g., MacGregor et al., 1986; Pierce et al., 1992). In these studies, menu alternatives were at first hidden and then exposed by the participant one-by-one whenever a down-arrow key was pressed. It is known that information acquisition behavior can be influenced by the cost of accessing it from the environment (Lohse & Johnson, 1996); therefore, it could be the case that the search behaviors observed by MacGregor et al.–self-terminating, exhaustive, and redundant–were a reflection of the cost structure imposed by the process-tracing methodology.

A solution to this problem is to infer a participant's search strategy from analysis of eye movement protocols. The active vision approach (Findlay & Gilchrist, 2003; Liversedge & Findlay, 2000) assumes that gaze shifts are tightly coupled with the allocation of visual perception and cognition. Eye movements provide an on-line indication of how people acquire and process information, and have provided significant benefits in the analysis of cognitive processes in a variety of task domains, such as reading (Just & Carpenter, 1984, 1980; Schilling, Rayner, & Chumbley, 1998), equation solving (Salvucci & Anderson, 2001), menu selection (Aaltonen, Hyrskykari, & Räihä, 1998; Byrne, Anderson, Douglass, & Matessa, 1999; Hornof, 2004), and web search (Card et al., 2001). Nonetheless caution is required when interpreting the assumed relationship between eye-movements and cognitive processes (see Anderson, Bothell, & Douglass, 2004).

In Experiment 1 eye movement protocols were recorded and from them we made inferences about participants' interactive search behavior. We adopt the convention found in the reading literature (e.g., Rayner & Pollatsek, 1989), where multiple successive fixations on an item are aggregated to an *item-gaze*. We refer to an item-gaze as a *visit*. We assume that an eye movement gaze directed towards an item in the menu can be broadly mapped to the cognitive process of making an assessment of the probability that an item will lead to the goal.

2.1. Method

Participants

Thirty-six Cardiff University undergraduate psychology students participated in return for course-related credit. All participants were native English speakers and had normal uncorrected vision. All participants were experienced in using a World Wide Web browser, and all had been required to use various computer software packages to produce coursework.

Design

Twelve critical (or experimental) trials were used to manipulate a single within-subjects factor, which was the relevance of the distractor items. In the critical trials, the

distractors had labeled links that were either of *moderate* or *poor* relevance to the goal statement. The label of the target item was always *highly-relevant* to the goal. The relevance of an item's label for a given goal was determined from ratings provided by a separate group of participants that did not take part in the experiment (see the next section for more details). Ratings of a labels' relevance were made on a 5-point scale, where five represented a label that was highly relevant and one represented a label that was of poor relevance to the goal description.

For each critical trial a single target item was selected that received a median relevance rating of five. The relevance of the remaining distractor items in the choice set varied between experimental conditions. In the moderate relevance distractor condition the median rating of the labels was three, and for the poor relevance distractor condition the median rating of the labels was one. The primary dependent measure was participants' eye movement data up to and including the first selection of an item.

In the critical trials the target item was always located towards the middle of the choice set. This was because if the target had been located towards the bottom of the menu, then the various hypotheses give identical behavioral predictions. So as to prevent participants learning that the target was always in the middle of the menu twelve filler trials were also used. In the filler trials the position of the target was random (excluding menu positions used in the critical trials).

The relevance of filler trial distractors was manipulated as a between-subjects factor: Half of the participants completed filler trials that contained moderate distractor items, and half of the participants completed filler trial that contained poor distractor items. The label of the target item was again always highly relevant to the goal. No further reporting of this manipulation is made. Data from the two conditions were pooled.

Menu materials and ratings

In order to derive ecologically determined goal statements a web usage survey was posted to under-graduate students in the School of Psychology at Cardiff University. The web usage survey aimed to identify typical search queries for the particular pool of participants. From the 25 responses to the survey (approximately 5% response rate) it was possible to determine 45 unique search goals. As an example two of the goal statements were "*Check your bank balance*" and "*Find a road map of Cardiff*".

The web usage survey also allowed us to discover which web sites the respondents had visited while searching the web for each goal. The labels from these suggested web sites were then sampled. For example, for the search goal "*check your bank balance*" labels were sampled from various online banking web sites (e.g., http://www.hsbc.co.uk and http://www.natwest.co.uk).

In order to put together menu choice sets for the experiment we had to first determine the relevance of each of the sampled web page labels in relation to a particular goal statement. A number of web-based tools are available to automatically compute the semantic similarity between a label and a goal description (e.g., LSA, Deerwester et al., 1990; Landauer & Dumais, 1997,available at http://lsa.colorado.edu/; GLSA, Royer, Farahat, Pirolli, Budiu, 2005, available at http://glsa.parc.com). Our initial attempts to gain estimates using LSA found that we either could not gain an estimate for many of the labels or that LSA provided many erroneous judgments of relevance. (Brumby (2005) reports details of the how LSA was used in order to gain estimates of label relevance, along with an appendix of the experimental materials used here.) LSA was not appropriate in this context because either an entry for the word did not exist in the training corpus or because the goal statement and labels were too brief for LSA to compute an accurate similarity score (but see Blackmon et al., 2003 for description of cases were LSA has been successfully used). Consequently, we opted to gain estimates of label relevance by gathering ratings from human judges.

In the ratings study, thirteen Cardiff University undergraduate psychology students completed a simple ratings questionnaire in return for course-related credit. (None of the participants that took part in the rating study took part in any of the subsequent experiments.) Participants were instructed to estimate the likelihood that selecting a label would lead to the achievement of the goal. Relevance estimates were made on a 5-point scale, where five represented a label that was very relevant to the goal description and one represented a label that was not at all relevant to the goal description. To gather the ratings all of the sampled web-labels for a particular goal description were made available at once and participants asked to rate them one-by-one. Participants were instructed to make each relevance estimate independently of the estimate of the relevance of the other labels presented on the page. Based on the ratings of the labels that were sampled from various web pages, it was possible to construct 16-item menus with labels of varying semantic relevance to a particular goal statement.

Procedure

In the experiment participants completed 24 search trials and 4 practice trials. There were six trials for each of the experimental conditions as well as 12 filler trials and four practice trials (i.e., 4 + 12 + 2 x 6). Each trial required the participant to search a simplified web page (or menu) for information relevant to a given goal statement. There was a different goal statement and set of labeled links for each trial; there was no systematic repetition of labels across trials. Each of the menu choice sets contained 16 labeled links, of which only one led to the completion of the goal (i.e., one target item and 15 distractor items). A purpose-built Microsoft Visual Basic program running on a PC with a high contrast 19-inch CRT monitor controlled the experiment. The items in each menu were presented in a standardized format: characters were font 16 Times New Roman and labels were presented in a single vertical list with an approximate distance

between each label of three degrees of visual angle. (Note: We found, in unreported pilot studies that such a large vertical separation between labels was required to gain an accurate mapping between eye movement fixations and items in the menu. It should be noted that in many cases web pages often present smaller fonts and tighter spacing than used here.) On filler trials the target was randomly positioned, and on critical trials, the target item was located between menu positions 3 and 8. As outlined above, this was in order that the eye-tracking data could clearly discriminate between the different theoretical accounts.

In the study, each trial commenced by the participant first reading the goal statement. When the participant was ready, they selected a search button with the mouse, which then presented the menu on the screen and removed from the screen the goal statement. Participants were instructed that for each trial in the experiment they would be required to select a label from the list of alternatives in the menu in pursuit of a given goal statement. More specifically, participants were instructed to select labels that they believed would be likely to lead to the goal. They were also informed that within each menu there was only one correct label and that the rest of the labels were distractors.

In order to impose a meaningful cost structure to the task, participants did not progress to the next trial until they correctly selected the single target item from each menu. When a selection was made the menu was removed from the screen and feedback on the accuracy of that selection was presented to the participant (i.e., labels read either "correct" or "incorrect"). If the participant did not correctly select the target item they were also told to press a back button in order to return to the menu and make an alternative selection from the menu. In other words, participants only progressed to the next trial in the experiment after selecting the actual target item. Participants were therefore encouraged to choose an item as quickly, but also as accurately as possible, in order to finish the study in a timely manner. Participants were free to leave as soon as they had completed the study.

Eye tracking was performed using an ASL Pan/Tilt optics eye tracking system. Eye movement data were sampled at a rate of 50 times per second (once every 20 ms). Eye movement fixations were determined using the Applied Science Laboratories *Eyenal* software package. Areas of interest were then defined as a standardized rectangular area around each menu item (occurring at the mid-point between vertically contiguous item, see Figure 1). Fixations were mapped to an item in the menu if they landed within the items respective area of interest. Fixations that did not land over a menu item were ignored (accounting for less than 4% of all fixations).

2.2. Results

For each trial we were only interested in participants' search behavior from the beginning of a trial up to the initial selection of an item. That is, data are not reported on search behavior that occurred after an incorrect selection, even though participants were

required to return to the menu and continue searching for the target item. For Experiment 1, all trials were analyzed regardless of correctness of participants' initial selection.

Figure 1 presents a typical eye movement fixation trace and a schematic representation of the collapsed gaze sequence from Experiment 1. In this example, the participant was given a goal statement to "Find a road map of Cardiff". During the experiment, only the labeled links were presented on the screen. The rectangular boxes around each menu item in Fig. 1a represent the areas of interest that were defined in order to map eye movement fixations to items in the menu. Fig. 1b presents a schematic representation of the gaze sequence. As we have said, we adopted the convention found in the reading literature (e.g., Rayner & Pollatsek, 1989) of collapsing multiple successive fixations on an item to a single *item-gaze*.

The gaze sequence in Fig. 1 provides an illustrative example of a number of interesting regularities observed in the eye movement data. Notice that the first few items in the menu were visited one after another in a top-to-bottom fashion (e.g., the first three visits in the gaze sequence were directed in turn to items 1, 2, and 3 in the menu). After visiting the target item, however, the participant decided not to select it immediately but continued to check some of the remaining items in the choice set. Not all items were visited in order as before, but instead the participant skipped over intermittent items while scanning down the list of options (e.g., after visiting item 3 the participant skips over a number of items in the middle of the menu before landing on item 11; further, all of the subsequent gaze transitions are also between non-neighboring items). This latter point will be returned to in more detail shortly.

(In what follows, we do not present an analysis of trial (or practice) effects. The reason for this is that the large and variable contribution of semantics to the behavior makes the effect of this variable extremely noisy. The experiment was not designed to examine practice effects.)

***** Figure 1 *****

Time required for and accuracy of initial selection

We conducted a set of analyses of the effects of semantic relevance on selection. A 2-tailed paired-samples t-test was used for these analyses. It was found that participants spent on average less time to make selection of an item when the distractors were of poor relevance (M = 5.47 s, SD = 1.71 s) compared to when the distractors were of moderate relevance to the goal statement (M = 7.58 s, SD = 3.52 s), t (35) = 5.16, $p < .001$. Participants were also more likely to accurately select the target on their initial selection when the distractors were of poor relevance (M = 98.15%, SD = 6.64%) compared to moderate relevance (M = 73.15%, SD = 16.07%), t (35) = 7.77, $p < .001$. Taken together these analyses indicate that participants were both quicker and more accurate in selecting

the target item when there was greater discrimination between target and distractor relevance. We next provide an analysis of eye movement protocols.

Number of items visited/revisited

The main eye-tracking measure of interest was the number of items that were visited at least once and also the number of items that were revisited (i.e., the number of items that were visited at least twice) for each experimental condition. These data are presented in Figure 2. It is apparent that participants visited on average approximately half of the items (8.56 items) in the menu at least once prior to selection. Revisits to items during search were also reasonably common. It can be seen that target discriminability affected search behavior: Participants visited and also revisited more of the items in the menu when the distractors were of moderate relevance to the target item than when they were of poor relevance. Two-tailed paired-samples t-tests found that the relevance of the distractor items had a significant effect on the number of items that were visited at least once, t (35) = 6.05, $p < .001$, and also the number of revisits to items that were made, t (35) = 5.57, $p < .001$. Moreover, participants decided to select an item without sampling all the items in the menu. The less relevant the distractors were, the fewer were visited.

***** Figure 2 *****

Proportion of first-visit-selections

There was evidence for two types of search strategy masked in the aggregate statistics reported above. Participants either chose to select an item after visiting for the first time or they continued to visit most, but not all, of the remaining items in the menu. In order to explore this hypothesis we considered the distribution of the frequency with which each number of items was visited after the initial visit to the selected item. That is, if the participant selected, for example, the fourth item in the menu, then we excluded all item visits that occurred before the initial visit to the fourth item in the menu (+/- 1 item) and counted the number of items that were visited at least once in the remaining gaze sequence.

Figure 3 shows the distribution of the number of items visited after the initial visit to the selected item. This distribution was clearly bimodal: There are two distributions one with a peak at 0 and the other at 9 items visited after the initial look at the selected item. This suggests that participants sometimes chose to select an item after visiting for the first time, a behavior we shall refer to as *first-visit-selection* (which is equivalent to MacGregor et al.'s (1986) description of self-terminating search). In addition, participants sometimes chose to continue to visit most, but not all, of the remaining items in the menu. This is reflected by the frequency difference in the latter distribution in Fig. 3, which probably reflects the difference in the position of the target item across trials.

***** Figure 3 *****

We investigated whether the relevance of the distractor items affected the likelihood that participants' chose to select an item immediately after visiting it for the first time. It was found that participants made more first-visit-selections on trials where the distractor items were of poor relevance to the goal (M = 36.57%, SD = 30.56%) than when they were of moderate relevance to the goal (M = 23.61%, SD = 22.32%), t (35) = 3.68, p < .001.

Duration of an item visit

A claim of the iterative deepening account is that participants should spend more time on each successive revisit to an item. Consequently, we considered whether the duration of item gaze increased between an initial visit and subsequent revisits. We also explored whether target discriminability affected the duration of item gazes. These data are presented in Figure 4. For statistical analysis of these data a 2 x 2 (distractor relevance x order of visit) repeated-measures ANOVA was used. We found that on average initial visits to an item were in fact longer in duration (M = 697.95 ms, SD = 204.62) than subsequent revisits to the same item (M = 436.80 ms, SD = 151.37), F (1, 35) = 32.28, p < .001, MSE = 76064.13. It was also found that the duration of all item gazes (both initial visits and revisits) were longer when the distractor items were of moderate relevance (M = 628.04 ms, SD = 130.76 ms) compared to when they were of poor relevance (M = 506.72 ms, SD = 121.58), F (1, 35) = 51.63, p < .001, MSE = 10263.28. There was not a significant interaction between distractor relevance and the order of visit on the duration of item gazes, F (1, 35) = 2.39, p = .131, MSE = 25083.25. This data suggests that participants looked at an items' label for longer when it was first encountered, and were willing to invest more time in looking at items that were of greater relevance to the goal.

***** Figure 4 *****

Further strategic adaptations

Analysis of the number of items fixated at least once suggested that people rarely fixated all of the items in the menu prior to selection (above). In fact on only 8.19% of searches were all 16 items in the menu visited prior to selection. Obviously, all of the items in the menu were unlikely to have been visited if an item was selected immediately after an initial visit. Analysis of eye movement protocols suggest that another reason that items in the menu were not visited was because participants frequently skipped over items as they scanned down to the bottom of the menu. Eye movement protocols show that when an item was fixated participants often did not fixate the next neighboring (or spatially contiguous) item in the list. Instead participants would often skip over the next spatially contiguous item and visit the second or third item in the list from the currently attended item. Furthermore, we observed that when participants chose not to select the target item after visiting it for the first time they continued assessing items in the menu,

but were more likely to skip over some of the intermediate items as they scanned down to the bottom of the menu.

We explored these observations by considering the probability that gaze transition occurred between non-contiguous items. We defined a skipping gaze transition as a gaze transition that did not occur between spatially contiguous (i.e., neighboring) items. The number of skipping gaze transitions was then divided by the total number of gaze transitions for a given trial (i.e., skipping-gaze-transitions / skipping-gaze-transitions + non-skipping-gaze-transitions). For this analysis we also excluded all upward gaze transitions (e.g., item 15 to item 4). This conservative analysis was adopted because we believed that most upward gaze transitions were motivated by the need to verify the location of an item for selection with the mouse, rather than by the need to make a new assessment. Furthermore, trials in which the total number of gaze transitions was less than or equal to 1 were excluded. Using the analysis defined above it was found that on average 42.49% (SD = 18.23%) of all downward gaze transitions were between non-neighboring items in the menu.

We explored the idea that the decision to skip over items may have been affected by whether a highly-relevant item had been encountered. For this analysis an additional factor was included that split the gaze sequence into transitions that occurred *before* and and *after* the initial fixation of the eventually selected item. Figure 5 shows these data split by distractor relevance. The figure shows that participants were more likely to make skipping gaze transitions after locating a candidate item for selection, but that this only happened when the distractors were of poor relevance. A 2 x 2 (distractor relevance x fixation of selected item) repeated-measures ANOVA was used for statistical analysis of these data. There was a significant main effect of whether or not the selected item had been fixated on the proportion of skipping gaze transitions, $F\,(1, 35) = 17.79$, $p < .001$, $MSE = .02$. The main effect of distractor relevance was not significant, $F\,(1, 35) = 3.66$, $p = .06$, $MSE = .02$. There was a significant interaction between whether the selected item had been fixated and distractor relevance on the proportion of skipping gaze transitions, $F\,(1, 35) = 4.23$, $p = .05$, $MSE = .02$. Further analysis of simple main effects found that the proportion of skipping gaze transitions significantly differed between whether or not the selected item had been fixated when the distractors were of poor relevance, $F\,(1, 35) = 26.72$, $p < .001$, $MSE = .02$, but not when distractors were of moderate relevance, $F\,(1, 35) = .73$, $p = .39$, $MSE = .02$. In other words, the decision to skip over items was affected by whether or not a candidate item for selection had been encountered. But this only happened when the previously visited distractors were of poor relevance to the goal.

***** Figure 5 *****

Finally, we observed that after first visiting the eventually selected item, participants would sometimes leave the mouse hovering over the item while they scanned over the remaining items in the menu. Interestingly, participants would then select the item with

the mouse without moving their eyes back to it (i.e., suggesting that the mouse was strategically left over the item to potentially minimize selection time, if no other competing item was found). This behavior occurred on approximately 16% of searches.

2.3. Discussion

Experiment 1 examined strategies for interactive search by investigating the consequences of manipulating distractor relevance for assessment and selection. Participants were found to have visited fewer of the available items in the menu when the distractors were less relevant (poor) to the goal statement. Whereas, when the distractors were more relevant (moderate) to the goal, more of the items in the menu were visited and also subsequently revisited prior to selection. These findings are contrary to the predictions of a simple threshold account, in which fewer assessments should be made given more relevant distractors: distractors of moderate relevance are more likely to exceed the threshold than distractors of poor relevance. The findings are consistent with rational accounts, in which it is assumed that the likelihood that the most recent item will lead to the goal is contingent on the estimated relevance of all assessed items and not just on the relevance of the most recently assessed item.

Further analysis indicated that participants were actually more likely to select an item immediately after visiting it for the first time (i.e., make a first-visit-selection) when the distractors were less relevant to the goal. These findings are consistent with accounts that assume that people adjust an estimate of the likelihood that a particular item will lead to the goal given the context provided by the likelihood that other items in the set will lead to the goal. The assessments that a participant makes of the item that they eventually select are not the only factor that determines selection; rather the likelihoods generated by these assessments need to be put in context of the assessments of alternative items. If a participant found many low relevance distractors in a choice set, then they may have judged other items to have a higher likelihood of success compared to when the distractors were highly relevant.

In addition, the results of Experiment 1 demonstrate that participants frequently revisit items while they search and were also more likely to do so when the distractors were of greater relevance to the goal. There are a number of possible explanations for why people might revisit items. First, a few item revisits would be expected due to the process of relocating an item for selection. Second, because human memories (or memory traces) are known to decay with time (Anderson & Schooler, 1991; Baddeley, 1990; Ebbinghaus, 1885), people might revisit relevant items in order to maintain them in memory (Peebles & Cheng, 2003). That is, people might return to items that they remember as being relevant to the goal. Third, revisits might reflect the application of different assessment methods. Further work is required to discriminate between these accounts.

Recall that a prediction of Rieman et al.'s (1996) iterative deepening account was that participants should spend more time on each successive revisit to an item. In the IDXL model each revisit reflected the use of increasingly high quality and costly assessment methods. The results of Experiment 1 do not support this prediction, however. Participants were in fact spending less, and not more, time dwelling on an item during revisits than on initial visits. This finding is inconsistent with the empirical observations of Rieman (1994) and Franzke (1994, 1995), a fact that may be due to the differences between tasks (i.e., learning to use a computer application vs. searching a web page for a goal-relevant label) or because of the different experimental methodology (i.e., verbal protocols vs. eye-tracking protocols). It is also possible that much less elaboration and reification of the meaning of labels is common during web search, than it is for people learning a complex computer application package. Either way, there is clearly more to be explained.

In addition to the main findings of Experiment 1, we also observed that participants sometimes skipped spatially contiguous items as they scanned down the list of items. Similar skipping behavior has previously been reported in studies of simple, routine menu selection (Aaltonen et al., 1998; Byrne et al., 1999; Hornof, 2004). In these studies, which were based on Nilsen's (1991) paradigm, participants were required to search a menu for a known target item, such as a single target letter or number amongst distractors. In contrast, interactive search tasks require the participant to estimate the probability that the selection of an option would lead to the goal based on the semantic match of the items label to the goal statement. It is unexpected that this label skipping behavior occurs during interactive search.

Furthermore, the results suggest that item skipping might be strategic. Participants were more likely to skip over items after they found a candidate item for selection; however, this only happened when the previously visited distractors were of poor relevance to the goal. When the distractors were of greater relevance spatially contiguous items were assessed in order. Participants may have been predicting that the set of unassessed items would be of similar quality to the set of assessed items. Taken together the results of Experiment 1 suggest that people are in fact more strategic and sensitive to context than previous models of interactive search suggest.

3. EXPERIMENT 2

In order to further investigate the context-sensitive account of interactive search we examined the consequences of manipulating the position of the target in the menu. The results of Experiment 1 suggest that the relevance of distractor items influenced whether people choose to select a highly relevant item immediately or choose to continue assessing items. Given these results we might expect the influence of distractor relevance on immediate selection to be diminished or even absent when a highly relevant item is encountered early in the menu because only a few of the distractors would have been assessed beforehand. In contrast, if a highly relevant item is encountered later, or near

the end of a menu, then many of the distractors will have been assessed beforehand and we should expect the full impact of distractor relevance on the decision to select an item immediately. In this circumstance people should be more likely to select immediately if distractors are less relevant and less likely to select immediately if distractors are more relevant. In other words the results of Experiment 1 should be replicated if the target is positioned late in the menu and be absent if it is positioned early in the menu.

In addition, the results of Experiment 1 suggest that the relevance of a highly relevant item should impact the assessment of subsequently encountered distractors. In context-sensitive accounts it is not only the relevance of distractors that affects the processing of targets but, in addition, target relevance should impact distractor processing. If a highly relevant item is encountered early in a menu then it provides a context for further assessments. Experiment 1 suggests that assessment may seem less beneficial when it is considered after an encounter with a highly relevant item than when it is consider prior to the selection of an item. One of the ways in which target item relevance can change assessment behavior is that people may be more likely to skip items. It follows that participants should exhibit more skipping behavior when a highly relevant item is encountered earlier in a menu compared to when it is encountered later in the menu.

3.1. Method

Participants

Sixteen Cardiff University undergraduate psychology students participated in return for course-related credit. None of the participants had previously taken part in the ratings study or Experiment 1. All participants were native English speakers and had normal uncorrected vision. All participants were experienced in using a World Wide Web browser and all had been required to use various computer software packages to produce coursework.

Design

The experiment manipulated the position of the target item and the relevance of the distractor items. There were two levels of target position (top part of the menu and bottom part) and three levels of distractor relevance (moderate, poor, and very poor). A within-subjects design was used. As in the previous study, each menu contained 16 labeled items. The manipulation of target position was counter-balanced across participants. For a given menu, the target was randomly positioned towards the top of the menu for half of the participants and was randomly positioned towards the bottom of the menu for the other half. Estimates of label relevance from the ratings study were used to devise menus (see Section 2.1 for more details on the ratings study). For each trial, a single target item was selected that received a median relevance rating of five. The relevance of the remaining distractor items in the choice set varied between experimental conditions. In the moderate relevance distractor condition the median rating of the labels

was three, for the poor relevance distractor condition the median rating of the labels was two, and for the poor relevance distractor condition the median rating of the labels was one. The primary focus of the study was on eye-tracking data of participants' eye movements up to and including the first selection of an item.

Materials and procedure

In the experiment participants completed 40 search trials. There were six trials for each of the experimental conditions, as well as four practice trials (i.e., 4 + 2 x 3 x 6). As in Experiment 1 each trial required the participant to search a simplified web page (or menu) for information relevant to a given goal statement. The goal statements were the same as those used in Experiment 1. There was a different goal statement and set of labeled links for each trial and there was no systematic repetition of labels across trials. Each of the menu choice sets contained 16 labeled links, of which only one led to the completion of the goal (i.e., one target item and 15 distractor items). For each trial the target item was placed in one of six random positions: three of these positions were towards the top half of the menu and three were in the bottom half of the menu.

As in Experiment 1, a purpose-built Microsoft Visual Basic program running on a PC with a high contrast 19-inch CRT monitor controlled the experiment. The items in each menu were presented in a standardized format: characters were font 16 Cosmic Sans MS and labels were presented in a single vertical list with an approximate distance between each label of three degrees of visual angle.

In the study, each trial commenced by the participant first reading the goal statement. When the participant was ready, they selected a search button with the mouse, which then presented the menu on the screen, and removed from the screen the goal statement. Participants were instructed that for each trial in the experiment, they would be required to select a label from the list of alternatives in the menu in pursuit of a given goal statement. More specifically, participants were instructed to select labels that they believed would be likely to lead to the goal. They were also informed that within each menu, there was only one correct label, and that the rest of the labels were distractors. Participants were instructed to commence their search at the top of the menu so as to ensure that the target was encountered either at an early or late stage in search. As before, in order to impose a meaningful cost structure to the task participants did not progress to the next trial until they selected the target item (i.e., the correct item). If they selected a distractor then they were presented with the same task again. This procedure was repeated until the target was correctly selected. Eye movement data were recorded using an ASL Pan/Tilt optics eye tracking system, which was sampled at a rate of 50 times per second. Eye movement fixations were determined using the same procedure outlined in Experiment 1.

3.2. Results

Accuracy

An analysis of selection accuracy data is presented. A 2 x 3 (target position x distractor relevance) repeated-measures ANOVA was used for statistical analysis. It was found that participants were less likely to accurately select the target item when the distractors were of moderate relevance to the goal (M = 79.17%, SD = 18.93%) compared to when they were of poor (M = 94.27%, SD = 10.03%) or very poor relevance to the goal (M = 96.88%, SD = 6.61%), F (2, 30) = 20.25, $p < .001$, MSE = .140. Selection accuracy did not differ between whether the target was located towards the top (M = 88.89%, SD = 7.86%) or the bottom of the menu (M = 91.31%, SD = 5.73%), F (1, 15) = .92, p = .35, MSE = .015. The position x distractor relevance interaction was also non-significant, F (2, 30) = 1.07, p = .36, MSE = .021.

For all subsequent analyses we only consider trials in which the participant correctly selected the target item on their initial selection. This differs from Experiment 1 where all trials were analyzed regardless of the accuracy of the initial selection. The reason for this difference was because in Experiment 2 the position of the target item was an independent variable. It was therefore important to exclude trials in which items other than the target were initially selected. Only 10% of trials were excluded from further analysis.

Time to selection

It was found that participants spent on average less time to select the target item when it was located towards the top of the menu (M = 6.52 s, SD = 2.93 s) compared to when it was located towards the bottom of the menu (M = 9.43 s, SD = 2.49 s), F (1, 15) = 28.267, $p < .001$, MSE = 7.179. The time to select the target item was greater when the distractors were of moderate relevance to the goal (M = 9.35 s, SD = 4.01 s), compared to when they were of poor (M = 7.15 s, SD = 2.13 s) or very poor relevance (M = 7.43 s, SD = 2.33 s), F (2, 30) = 17.82, $p < .001$, MSE = 2.57. The position x distractor relevance interaction was not significant, F (2, 30) = .12, p = .89, MSE = 5.08.

Proportion of first-visit selections

We investigated whether the position of the target and relevance of the distractors affected the likelihood that participants chose to select an item immediately after visiting it for the first time. It can be seen in Figure 6 that participants were more likely to commit to a first-visit-selection when the target item was positioned towards the bottom of the menu, than when it was positioned towards the top of the menu. It can also be seen that participants were more likely to make a first-visit-selection on trials where the distractor were of less relevance to the goal, compared to when they were of moderate relevance. A 2 x 3 (target position x distractor relevance) repeated-measures ANOVA found a significant main effect of target position on the percentage of trials that a first-visit-selection was made, F (1, 15) = 18.05, $p < .001$, MSE = .07. There was also a

significant main effect of distractor relevance, F (2, 30) = 6.76, p = .006, MSE = .03. The position x distractor relevance interaction was not significant, F (2, 30) = 2.23, p = .125, MSE = .06.

***** Figure 6 *****

The absence of statistically significant interaction is disappointing given that the context-sensitive account predicts the presence of an effect of distractor relevance only when the target is positioned toward the end of the menu. Nonetheless we conducted planned comparisons of the simple effect of distractor relevance at each level of target position.

As the data in Fig. 6 suggest the effect of distractor relevance was contingent on the position of the target. When the target was located towards the bottom of the menu participants were significantly more likely to select an item immediately after visiting it for the first time when the distractors were less relevant to the goal, F (2, 14) = 7.81, p = .005. But when the target was located towards the top of the menu, there was no effect of distractor relevance on the percentage of trials on which a first-visit-selection was made, F (2, 14) = .123, p = .885.

Skipping gaze transitions during interactive search

Next we investigated whether the position of the target affected participants' propensity for skipping over some of the items in the menu. Recall that it was predicted that on average the number of gaze transitions between non-neighboring items would be greater for trials where the target item was positioned towards the top of the menu compared to the bottom. This prediction was derived from the hypothesis that participants would be less likely to visit every item in turn after visiting, but not necessarily selecting, the target item.

As in Experiment 1, we defined a skipping gaze transition as a gaze transition that did not occur between spatially contiguous items. For each trial, the number of skipping gaze transitions was divided by the total number of gaze transitions. Trials in which the number of gaze transitions were less than or equal to 1 were excluded. For this analysis we also excluded all upward gaze transitions. This amounted to 23.07% (SD = 16.43%) of all gaze transitions across participants being excluded because they traveled in an upward direction.

Data were analyzed to investigate whether there was an effect of target position on the proportion of skipping gaze transitions. It was found that proportionally more gaze transitions were between non-neighboring items when the target item was located towards the top of the menu (M = 54.12%, SD = 7.89%) compared to when it was located towards the bottom of the menu (M = 49.45%, SD = 6.96%), F (1, 15) = 10.86, p < .005, MSE = .005).

We also considered whether the relevance of the distractor items affected the proportion of skipping gaze transitions. Although participants were slightly more likely to skip over items when the distractors were of moderate relevance (M = 55.31%, SD = 1.27%) compared to when they were of poor (M = 50.15%, SD = 6.80%) or very poor relevance to the goal (M = 50.18%, SD = 7.01%), analysis found that this effect was not statistically significant, F (2, 30) = 1.92, p = .16, MSE = .001. It is worth noting that this lack of main effect of distractor relevance on the proportion of skipping gaze transitions is consistent with the results in Experiment 1. The position x distractor relevance interaction was also non-significant, F (2, 30) = .08, p = .92, MSE = .001.

Frequency of visits to each item location

While the target items were randomly positioned in the menu, the positions were biased toward the top and the bottom of the menu in accordance with the experimental design. It was possible, therefore, that the participants would learn to examine the top and bottom of each menu without looking at items in the middle. Such a bias would be problematic because it would suggest that people were navigating by guessing the location of the target item in the menu, rather than by following scent. To test whether this occurred we examined the frequency that each item location in the menu was visited at least once.

Figure 7 shows the frequency with which each item location in the menu was visited at least once. When the target was located towards the top of the menu then there was a step-like function in the distribution. This distribution reflects the difference in trials where participants chose to select the target immediately (i.e., make a first-visit-selection) and where they chose to make further assessments. When the target was at the bottom of the menu there was a flatter distribution of visits. There is no evidence in either distribution that participants preferred to visit top or bottom items at the expense of middle items.

***** Figure 7 *****

3.3. Discussion

With Experiment 2 we manipulated the position of the target item within the choice set in order to further test the hypothesis that the decision of when to select an item is dependent on the set of assessments previously made. The results show that if the target had been positioned towards the bottom of the menu participants were more likely to select it immediately when the previously visited distractors were less relevant to the goal. In contrast, when the target was positioned towards the top of the menu the relevance of the distractors had no measurable effect on the decision to select it immediately. These findings suggest that the greater the number of distractors assessed the greater the effect of their relevance on immediate selection.

In addition, the results of Experiment 2 demonstrated that whether the target item had been encountered had an effect on the manner in which further distractors were assessed. When the target was encountered early in the search process participants were more likely to skip items (i.e., they would make more non-contiguous gaze shifts). The findings suggest that human interactive search is not only sensitive to the context provided by low relevance distractors, but also by the context provided by highly relevant, but as yet unselected items. Although people may not commit to an immediate selection of a highly relevant item, the assessment of such an item does have consequences for subsequent assessments. To the best of our knowledge, none of the current models of interactive search provide a direct explanation for this pattern of item skipping behavior.

The main findings of Experiment 2 help distinguish between alternative hypotheses concerning the strategies that people use for interactive search. In particular, because it was found that the decision to select the target was shown to be dependent on its position within the set of options, these findings do not support the hypothesis that people assess until the most recently assessed item exceeds a threshold.

The main finding of the current experiment is consistent with the predictions of the ACT-R model proposed by Brumby and Howes' (2004). Brumby and Howes proposed that an item became a candidate for selection once the activation of the item's representation in memory had exceeded a threshold. However, the activation of an item was determined by its relevance to the goal and, through attentional focusing, by the relevance of other items in the set. It was therefore the case that, for example, a negative assessment of an item y could lead directly to an increase in the activation of a previously assessed item x. The model predicted that participants would be less likely to select an item without further assessment if that item occurred earlier rather than later in the choice set. Selection without further assessment would be even less likely if distractor items were less relevant to the goal. These predictions differ from a similar context sensitive account of interactive search proposed by Cox and Young (2004).

Cox and Young's model makes the prediction that participants would be more likely to select an item without further assessment if that item occurred earlier, rather than later, in the choice set. The data show the opposite pattern. The prediction from Cox and Young's model was due to the assumption that when a goal-relevant item was encountered very early on then the most efficient thing to do would be to invest further in that item by iteratively performing a more costly but higher quality assessment of that item (which would in turn lead to selection of the item). In contrast the data suggest that when a goal-relevant item is encountered, instead of investing further in that item, people opt to continue assessing the remaining items in the menu with a low cost, low benefit, assessment (perhaps making use of the time required to move the mouse to the target and select the item). In contrast, the findings do not support one particular prediction made by Cox and Young (2004).

4. GENERAL DISCUSSION

Two experiments were used to systematically manipulate the relevance of distractor items and the location of the target item within menus. Eye movement fixations, between onset and selection, were used to calculate the number of visits made to each item and in turn infer which items were assessed. In Experiment 1, manipulation of distractor relevance had consequences for when participants decided to select an item: participants were more likely to select a highly-relevant item, without looking at any further items, when the distractors were less relevant to the goal. In Experiment 2, immediate selection of the target was more likely when more of the items in the choice set had already been assessed, especially if those items were less relevant to the goal. In addition, skipping was more likely with early exposure to a highly relevant item.

The experiments provide evidence to support the view that people are sensitive to the context provided by previous assessments when deciding whether to continue to assess items or to make a selection. By doing so, they also support a rational view of when people commit to a selection in a single menu page (e.g., Brumby & Howes, 2004; Cox & Young, 2004; Young, 1998). Participants did not make an a-priori commitment to either assess all of the items or to assess items until the value of the most recently assessed item was above a threshold (what we call a simple threshold account). The eye-tracking data showed that people often did not visit all of the items in the menu and provided no evidence to support a simple threshold account. According to a simple threshold account, when distractors are more relevant people should look at fewer items because it is more likely that one of the distractors will be above threshold. Data from both of the reported studies indicate that this is not the case, however, because participants in the studies tended to visit more items when the distractors were more relevant to the goal statement.

Consistent with the idea that human interactive search is rational, the experiments demonstrate that people may be sensitive to the expected value of conducting further assessments. When assessments were relatively valuable, people tended to conduct more assessments. For example, (1) when distractors were highly relevant (Experiment's 1 and 2) and there was therefore greater chance of error, people conducted more assessments; (2) when fewer items had already been assessed, fewer items had been rejected, and people were therefore less likely to select a candidate target immediately (Experiment 2), preferring instead to continue assessment; (3) once a candidate target had been identified, people conducted lower cost assessments (i.e., they assessed the remaining items, but were more likely to skip over some of them).

The findings were consistent with the ACT-R model reported by Brumby and Howes (2004), which was described earlier in the current article. Following Young (1998), Brumby and Howes assumed that an accurate model of interactive search should be sensitive to the cross-dependencies between the likelihoods that items within a single menu set would lead to the goal. For example, a reduction in the estimate of the

likelihood that one item will lead to the goal should increase the estimate of the likelihood that another will do so. Brumby and Howes modeled these cross-dependencies with ACT-R's attentional focusing mechanism. A fixed amount of source activation was distributed between declarative representations of the menu items: The more and stronger the links between the goal statement and the representation of an item in declarative memory the higher that representations' activation. Assessments could result in the strengthening/weakening of associations between the goal and the representation. An assessment suggesting that one item was weak would lead to the redistribution of what was its activation to the representation of other items in memory. If the activation of a representation was sufficiently high, then it would become a candidate for selection. Importantly, the way in which the fixed source activation was redistributed, following each assessment, meant that it was not necessarily the most recent item that would become a candidate for selection.

In addition to the main findings concerning the effects of distractor relevance, we also found that participants frequently revisited items prior to selection and that items that were more relevant to the goal were more likely to be revisited. This finding partially supports the idea that participants were exhibiting iterative deepening of attention (Rieman et al., 1996; Young, 1998); however, we found that the duration of revisits to an item were on average shorter than earlier visits. This latter finding does not support the idea that when a goal-relevant item is encountered people invest further in that item by performing a more costly, but higher quality, assessment of the item.

While we did not find evidence that people revisit items to assess them with increasingly costly assessment methods,` we did find evidence that suggested that participants were adopting more than one kind of assessment procedure. In both of the studies, participants frequently skipped items in the menu. In particular, Experiment 2 found that when the target was encountered early in the search process people were more likely to skip items (i.e., they would make more non-contiguous gaze shifts).

In the remainder of the General Discussion, we consider three issues: (1) The interpretation of the skipping behavior; (2) issues concerning measures of relevance; (3) ecological validity.

4.1. The Interpretation of Item-Skipping Behavior

An interpretation of the skipping behavior is that it reflects the use of a low quality, low cost assessment method during interactive search. From this perspective, it is assumed that people make choices between different assessment methods that vary in their costs and potential benefits. Implicit in this account is the idea that people are rapidly assessing multiple items within a single eye movement. Given that items in the menu were separated by a large degree of visual angle, it may be the case that even low-level visual information about multiple items could be accessed within a single eye movement. To address this question, we briefly discuss different theories regarding the

range of the human effective field of view (or perceptual span) that defines the region from which the visual perceptual system processes information in a single fixation.

Cognitive architectures (e.g., EPIC: Kieras, & Meyer, 1997 and ACT-R: Anderson et al., 2004) commit to different theoretical assumptions regarding the processing limitations of the human visual attention system, and therefore do not provide a single answer as to whether participants could have assessed multiple items in the menu within one eye movement. Although the ACT-R theory makes no commitments to constraints on the degree of visual angle through which visual attention can shift, a theory has been proposed concerning the interaction between visual attention and eye movements (Salvucci, 2001). Salvucci (2001) proposed a model influenced by models of eye-movement control in reading, particularly the E-Z Reader model (Reichle, Rayner, & Pollatsek, 2003; Reichle, Pollatsek, Fisher, & Rayner, 1998). The theory assumes that shifts of *visual attention* underlie eye movements: Eye movements are a response to shifts of visual attention and are prepared and executed whenever the eye movement processor becomes available after the previous eye movement. Importantly, the theory assumes that shifts of visual attention can occur before the eyes move to the next item in the menu. Therefore, the next item in the menu is processed parafoveally while the eyes are fixated on the previous item and while an eye movement is being prepared. If that processing leads to the rejection of that item, the next item, then eye movement programming can be redirected to the item after the next item. This account would predict that if items were assessed with a low time cost assessment procedure, then, although every item would be assessed, the eyes would only be required to move every two or three shifts of attention, and thus skip over items. From Salvucci's (2001) perspective it would therefore seem plausible that multiple items could be assessed within a single eye movement, regardless of the large degree of visual angle between items in the menu.

In contrast to ACT-R, the EPIC (Meyer & Kieras, 1997) architecture is committed to constraints on the relationship between visual processing and visual angle. Given these constraints, Hornof (2004) proposed an EPIC model that implemented a maximally efficient foveal sweep strategy. In Hornof's model, visual information about multiple items was available simultaneously, but only if they fell within the foveal region. However, as the foveal region was constrained to be one degree of visual angle, it would seem that only one row of text could be processed within a fixation. From this perspective it would not be plausible to assume that multiple menu items could be assessed within a single eye movement, because at three degrees of visual angle (as used in our experiments) the distance between items in the menu was too great. Further work is required to discriminate between these accounts, because it is unclear which provides a better model of the data reported in the current article.

In addition, there is at least some empirical data to support the idea that the participants in the reported experiments could have assessed multiple items in the menu

within a single eye movement. A study by Ojanpää, Näsänen, and Kojo (2002) investigated the span of the effective visual field in vertical lists of words. The span of the effective visual field (McConkie & Rayner, 1975) refers to the area of the visual field in which words can be identified. The results suggest that the field extends 2.4–3 degrees of visual angle in a vertical direction from the point of fixation. This meant that participants in Ojanpää et al.'s study were able to identify on average up to four items from a vertically arranged menu during a single fixation. The number of items identified was affected by the size of the vertical separation between items: The larger the spacing the fewer the number of items were identified. The fact that the span of the effective visual field extends up to 3 degrees suggests, in the context of the current discussion, that participants might have been able to identify multiple items during a single eye movement fixation.

4.2. Issues Concerning Measures of Relevance

A methodological issue with the current study was the measure of relevance that was used to devise the various experimental conditions. In the ratings study, estimates of relevance for a each member of a set of labels were made after the presentation of a single goal (i.e., all labels for a particular goal statement were presented at the same time). Therefore, it is possible that the other labels available influenced label ratings. As a consequence it is not clear that the label ratings used in the experiment were independent. Participants may, for example, have made finer ranking decisions than they would have otherwise done. A better method for gather ratings may have been to present individual pairs of *<goal statement> : <label>* in a randomized order across participants. Such a methodology would perhaps have had the benefit of reducing, to some degree, the influence of any one ranking on another. This method creates its own problems, however, namely maintaining participant motivation to provide accurate estimate during what would be a rather lengthy and tedious task.

Alternatively, semantic systems might have been used to provide an absolute scale for determining mutual similarity estimates between the labels and the goal statements (e.g., LSA, Deerwester et al., 1990; Landauer & Dumais, 1997,available at http://lsa.colorado.edu/; GLSA, Royer et al., 2005, available at http://glsa.parc.com). While these techniques have been shown to reflect human performance quite accurately over very large data sets (Landauer & Dumais, 1997), the reported experiments used a relatively small sample of labels (16 labels per trial). In our experience these techniques are not sensitive enough to provide accurate estimates of relevance on such a small sample. Indeed, we found in an earlier study (Brumby, 2005) that LSA provided many erroneous judgments of label relevance compared to human judgments.

4.3. Ecological Validity of Menus

One concern is that while we are interested in how people search for information using the World-Wide Web, we have only looked at one particular type of task applied to a restricted class of site designs. A particular concern for generalizing the results reported here to how people search real web pages is that items in the experimental menus were separated by a high degree of visual angle. In many cases web pages appear to use smaller fonts and tighter spacing than used in the studies reported above. The reason for adopting a large vertical separation between labels was to gain an accurate mapping between eye movement fixations and items in the menu. It is possible that the spacing caused subjects to adopt a different strategy than they would on a real web page.

It is known that the time to locate a simple target item (e.g., a single character or word) in a visual search task is affected by the density of items in the display. Search time can either decrease as the density of items increases (Bertera & Rayner, 2000; Ojanpää, et al., 2002) or increase as the density of items increases (Halverson & Hornof, 2004a, 2004b) depending on the way in which density is manipulated. Also, it is generally assumed that the density of items in a display affects the number of items that can be perceived in a single fixation (see discussion above regarding effective field of view), which might also affect search time. Everett and Byrne (2004) have also found that density affects people's search strategy; although this study did not involve text-based materials, but rather participants searched a display for a target icon amongst distractor icons. Further work is needed to understand the implications of spacing for the choices that people make about what to fixate and when to select.

Another area that requires empirical investigation is whether the current findings replicate in interactive search tasks that takes place on small screen devices, such as a cellular phone or a PDA. Cellular phones offer a range of functionality other than simply making calls (including tools for managing contact information, voice mail, hardware settings, and often software for playing games and browsing the Web) that is often accessed through a menu structure. As discussed above, the tighter spacing between items on such mobile devices might have consequences for people's search strategy. Furthermore, in terms of the classic depth vs. breadth trade-off, there is at least some empirical evidence to suggest that a broad navigation structure, which as discussed previously was found to be superior during search on personal computers, also has an advantage for a small-screen device such as the cellular phone (Parush & Yuviler-Gavish, 2003). This finding is interesting because cellular phones generally have greater interaction costs than traditional personal computers (e.g., cellular phones requiring button presses for navigation and selection actions whereas personal computers support mouse movements). It is known that information acquisition behavior is influenced by the cost of accessing information from the environment (Lohse & Johnson, 1996). Further empirical work is required to determine whether strategies for exploring the menu structure on a cellular phone differ from those on a personal computer that have lower interaction costs.

Finally, it is also worth commenting on the relationship of our work to more general web-based activities, such as searching the results list of a search engine (e.g., Google, MSN Search, & Yahoo). Search engines are of particular interest at the moment because they provide a powerful tool to support a user's goal-directed search of the web. It is worth considering a number of important differences between the results page of a search engine and the type of interactive search task described here. For instance, search engines typically list results in order of their relevance to the query term, whereas the type of interactive search task used here assumed an unordered list of results. The results of a recent eye-tracking study (Granka, Joachims, & Gay, 2004), which examined how users interact with the results page of a search engine, found that users focused their attention on the top few (most relevant) items and rarely assessed all of the results prior to selecting an item.

5. CONCLUSION

The question articulated at the start of this paper was: How do people search a newly encountered web page for a link that is relevant to the achievement of their search goal? Estimates of label relevance are clearly important, but strategy also plays a role in guiding search. The empirical studies reported in this paper were aimed at providing evidence to support the view that rather than adopt a simple heuristic strategy (e.g., Assess-all or satisfice), people rationally adjust their judgment of the benefit of further assessments in light of the relevance of the set of items that have already been assessed. The reported studies systematically manipulated the relevance of the distractor items and the location of the target item within the set. People rarely visited all items prior to selection. More importantly, when the relevance of distractors was relatively low, a high value item was more likely to be selected when it was first visited. Conversely, when the relevance of distractors was relatively high, participants were less likely to select the target when it was first visited; they preferred instead to make more assessments of more items. A consequence of more highly relevant distractors was therefore that more items were visited prior to selection; a finding that is inconsistent with a satisficing account of interactive search.

In addition, the location manipulation (Experiment 2) suggested that if more distractors were assessed prior to an encounter with the target item, then the relevance of the distractors had a greater influence on the decision as to whether to select the target immediately. It was also found that an early encounter with a target item caused participants to increase the frequency with which they skipped visits to menu items, perhaps because they were adopting a lower cost assessment strategy.

Together the findings support the hypothesis that people adopt a rational strategy in deciding whether to continue to assess or to make a selection when searching a newly encountered web page. The findings support the rational view because they support the assumption that it is rational to adjust likelihood estimates given the context provided by the set of relevance estimates already made. Further, the findings demonstrate that

people may be sensitive to the expected value of conducting further assessments. When assessments were likely to be relatively valuable, people tended to conduct more assessments. In contrast, when assessments were less likely to be valuable, e.g. once a candidate target had been identified, people conducted lower cost assessments (in particular, they adopted a visual skipping strategy).

NOTES

Acknowledgments. We gratefully thank Richard Young and Anna Cox for their collaboration over the course of the development of the work presented in this paper, and also for providing us with the source code for their model. We would also like to thank Richard Young, Mike Byrne, Roger Remington and an anonymous reviewer for their many thorough and insightful comments on earlier drafts of this manuscript.

Support. Duncan Brumby was supported by an EPSRC school studentship given to the School of Psychology at Cardiff University.

Authors' Present Addresses. Duncan Brumby is now at Department of Computer Science, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. Andrew Howes is now at the Manchester Business School, University of Manchester, Booth Street West, Manchester, M15 6PB, UK.

HCI Editorial Record. (supplied by Editor)

REFERENCES

Aaltonen, A., Hyrskykari, A., & Räihä, K.-J. (1998). 101 spots, or how do users read menus? In C.-M. Karat, A. Lund, J. Coutaz, & J. Karat (Eds.), *Proceedings of the ACM CHI 1998 Human Factors in Computing Systems Conference* (pp. 132–139). New York, NY: ACM Press.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillside, NJ: Lawrence Erlbaum.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 5*, 451–474.

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval. *Psychological Science, 15*(*4)*, 225–231.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(*4)*, 1036–1060.

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

Anderson, J. R., & Pirolli, P. L. (1984). Spread of activation. *Journal of Experiment Psychology: Learning, Memory, and Cognition, 10(4)*, 791–798.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2(6)*, 396–408.

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Hillside, NJ: Lawrence Erlbaum.

Bertera, J.H., & Rayner, K. (2000). Eye movements and the span of effective stimulus in visual search. *Perception & Psychophysics, 62(3)*, 576–585.

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In G. C. van der Veer, & C. Gale (Eds.), *Proceedings of the ACM CHI 2005 Human Factors in Computing Systems Conference* (pp. 31–40). New York, NY: ACM Press.

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2003). Repairing usability problems identified by the cognitive walkthrough for the web. In G. Cockton, & P.

Korhonen (Eds.), *Proceedings of the ACM CHI 2003 Human Factors in Computing Systems Conference* (pp. 497–504). New York, NY: ACM Press.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. In L. Terveen (Ed.), *Proceedings of the ACM CHI 2002 Human Factors in Computing Systems Conference* (pp. 463–470). New York, NY: ACM Press.

Brumby, D. P. (2005). *An empirical investigation into strategies for guiding interactive search*. Unpublished doctoral dissertation, Cardiff University, Cardiff, UK.

Brumby, D. P. & Howes, A. (2004). Good enough but I'll just check: web page search as attentional refocusing. In M. Lovett, C. Schunn, & P. Munro (Eds.), Proceedings of the *6th Internal Conference on Cognitive Modeling* (pp. 46–50), Mahwah, NJ: Lawrence Erlbaum.

Byrne, M. D., Anderson, J. R., Douglass, S., & Matessa, M. (1999). Eye tracking the visual search of click-down menus. In M. W. Altom, & M. G. Williams (Eds.), *Proceedings of the ACM CHI 1999 Human Factors in Computing Systems Conference* (pp. 402–409). New York, NY: ACM Press.

Byrne, M. D., John, B. E., Wehrle, N. S., & Crow, D. C. (1999). The tangled web we wove: A taskonomy of www use. In M. W. Altom, & M. G. Williams (Eds.), *Proceedings of the ACM CHI 1999 Human Factors in Computing Systems Conference* (pp. 544–551). New York, NY: ACM Press.

Card, S. K., Pirolli, P., Van Der Wege, M., Morrision, J. B., Reeder, R. W., Schraedley, P. K., & Boshart, J. (2001). Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability. In M. Beaudouin-Lafon, & R. J. K. Jacob (Eds.), *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference* (pp. 498–505). New York, NY: ACM Press.

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the world wide web. In P. H. Enslow, & D. Kroemker (Eds.), *Proceedings of the Third International World-Wide Web Conference on Technology, Tools and Applications (*pp. 1065–1073). New York, NK: Elsevier.

Chi, E. H., Rosien, A., Suppattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., & Cousins, S. (2003). The Bloodhound project: Automating discovery of web usability issues using the InfoScent simulator. In G. Cockton, & P. Korhonen (Eds.), *Proceedings of the ACM CHI 2003 Human Factors in Computing Systems Conference* (pp. 505–512). New York, NY: ACM Press.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions on the web. In M. Beaudouin-Lafon, & R. J.

K. Jacob (Eds.), *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference* (pp. 490–497). New York, NY: ACM Press.

Chi, E. H., Pirolli, P., & Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In T. Turner, G. Szwillus, M. Czerwinski, F. Peterno, & S. Pemberton (Eds.), *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference* (pp. 161–176). New York, NY: ACM Press.

Cockburn, A., & McKenzie, B. (2001). What do web users do? An empirical analysis of web use. *International Journal of Human-Computer Studies, 54*, 903–922.

Cox, A.L., & Young, R.M. (2004). A rational model of the effect of information scent on the exploration of menus. Poster session at the meeting of the *6th Internal Conference on Cognitive Modelling*, Pittsburgh, PA.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6)*, 391–407.

Ebbinghaus, H. (1885). *Ueber das Gedachtnis*. Leipzig: Dunker (Translation by H. Ruyer and C.E. Bussenius (1913), *Memory*. New York: Teachers College, Columbia University.)

Everett, S.P., & Byrne, M.D. (2004). Unintended effects: varying icon spacing changes users' visual search strategy. In E. Dykstra-Erickson & K.M. Tscheligi (Eds.), *Proceedings of the ACM CHI 2004 Human Factors in Computing Systems Conference* (pp. 695–702). New York, NY: ACM Press.

Farahat, A., Pirolli, P., Markova, P. (2004). Incremental methods for computing word pair similarity (TR-04-6-2004). Palo Alto, CA: Palo Alto Research Center Incorporated.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: the psychology of looking and seeing.* Oxford, UK: Oxford University Press.

Franzke, M. (1995). Turning research into practice: Characteristics of display-based interaction. In I. R. Katz, R. L. Mack, L. Marks, M. B. Rosson, & J. Nielson (Eds.), *Proceedings of the ACM CHI 1995 Human Factors in Computing Systems Conference* (pp. 421–428). New York, NY: ACM Press.

Franzke, M. (1994). Exploration and experienced performance with display-based systems *(*Doctoral Dissertation, University of Colorado, 1994). *Dissertation Abstracts International: Section B: The Sciences and Engineering, 56*(2-B), 1134.

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. Poster session presented at the *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Sheffield, UK.

Halverson, T., & Hornof, A.J. (2004a). Local density guides visual search: Sparse groups are first and faster. *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society* (1860–1864). Santa Monica, CA: Human Factors and Ergonomics Society.

Halverson, T., & Hornof, A.J. (2004b). Strategy shifts in mixed-density search. In K. D. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 529–534)*,* Mahwah, NJ: Lawrence Erlbaum.

Halverson, T., & Hornof, A. J. (2004). Link colors guide a search. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1367–1370)*,* Vienna, Austria. ACM Press, New York, NY.

Hornof, A.J. (2004). Cognitive strategies for the visual search of hierarchical computer displays. *Human-Computer Interaction, 19*, 183–223.

Howes, A. (1994). A model of the acquisition of menu knowledge by exploration. In B. Adelson, S. Dumais, J. S. Olson (Eds.), *Proceedings of the ACM CHI 1994 Human Factors in Computing Systems Conference* (pp. 445–451). New York, NY: ACM Press.

Howes, A., Payne, S. J., & Richardson, J. (2002). An instance-based model of the effect of previous choices on the control of interactive search. In W. D. Gray, & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 476–481)*,* Mahwah, NJ: Lawrence Erlbaum.

Just, M. A., & Carpenter, P. A. (1984). Using eye fixations to study reading comprehension. In D. E. Kieras & M. A. Just (Eds.), *New Methods in Reading Comprehension Research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review, 87*, 329–354.

Katz, M. A., & Byrne, M. D. (2002). Effects of scent and breadth on use of site-specific search on e-commerce web sites. *ACM Transactions on Computer-Human Interaction, 10*(*3)*, 198–220.

Kaur, I., & Hornof, A. J. (2005). A comparison of LSA, wordNet and PMI-IR for predicting user click behavior. In G. Veer, & C. Gale (Eds.), *Proceedings of the*

ACM CHI 2005 Human Factors in Computing Systems Conference (pp. 51–60). New York, NY: ACM Press.

Kieras, D.E., & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12,* 391–438.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(*2)*, 211–240.

Larson, K., & Czerwinski, M. (1998). Web page design: Implications of memory, structure and scent for information retrieval. In C.-M. Karat, A. Lund, J. Coutaz, & J. Karat (Eds.), *Proceedings of the ACM CHI 1998 Human Factors in Computing Systems Conference* (pp. 25–32). New York, NY: ACM Press.

Lee, E., & MacGregor, J. (1985). Minimizing user search time in menu retrieval systems. *Human Factors, 27*(*2)*, 157–162.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Science, 4*(*1)*, 6–14.

Lohse, G. L., & Johnson, E. J. (1996). A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes, 68*(*1)*, 28–43.

MacGregor, J., Lee, E., & Lam, N. (1986). Optimizing the structure of database menu indexes: a decision model of menu search. *Human Factors, 28*(*4),* 387–399.

McCarthy, J. D., Sasse, M. A., & Riegelsberger, J. (2003). Could I have the menu please? An eye tracking study of design conventions. In E. O'Neill, P. Palanque, & P. Johnston (Eds.), *People and Computers XVII – Designing for Society* (pp. 401–414). London, UK: Springer-Verlag.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics, 17*, 578–586.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: implications for information architecture. *Human-Computer Interaction, 19*, 225–271.

Morrison, J. B., Pirolli, P., & Card, S. K. (2001). A taxonomic analysis of what world wide web activities significantly impact people's decisions and actions. Poster session presented at *CHI 2001 Extended Abstracts on Human Factors in Computing Systems Conference*, Seattle, WA.

Nilsen, E. L. (1991). Perceptual-motor control in human-computer interaction. Unpublished technical report, University of Michigan, Ann Arbor, MI.

Norman, K. L. (1991). *The psychology of menu selection: designing cognitive control of the human/computer interface.* Norwood, HJ: Ablex.

Ojanpää, H., Näsänen, R., & Kojo, I. (2002). Eye movements in the visual search of word lists. *Vision Research, 42*, 1499–1512.

Parush, A., & Yuviler-Gavish, N. (2004). Web navigation structures in cellular phones: The depth/breadth trade-off issue. International *Journal of Human Computer Studies, 60* (*5-6)*, 753–770.

Payne, S.J., Richardson, J., & Howes, A. (2000). Strategic use of familiarity in display-based problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26(6),* 1685–1701.

Pearson, R., & van Schaik, P. (2003). The effect of spatial layout of and link color in web pages on performance in a visual search task and interactive search task. *International Journal of Human-Computer Interaction, 59*, 327–353.

Peebles, D., & Cheng, P. C.-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors, 45(1),* 28–45.

Pierce, B. J., Parkinson, S. R., & Sisson, N. (1992). Effects of semantic similarity, omission probability and number of alternatives in computer menu search. *International Journal of Man-Machine Studies, 37*, 653–677.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science, 29(3)*, 343–373.

Pirolli, P., & Fu, W-T.F. (2003). SNIF-ACT: A model of information foraging on the world wide web. In P. Brusilovsky, A. Corbett, & F. de Rosis (Eds.), *User Modeling 2003, 9th International Conference on User Modelling (*Vol. 2702, pp. 45-54)*,* London, UK: Springer-Verlag

Pirolli, P., & Card, S.K. (1999). Information foraging. *Psychological Review, 106,* 643–675.

Rayner, K. & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hill.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*, 445–526.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*, 125–157.

Rieman, J. (1994). Learning strategies and exploratory behavior of interactive computer users (Doctoral dissertation, University of Colorado). *Dissertation Abstracts International: Section B: The Sciences and Engineering, 55*(10-B), 4470.

Rieman, J., Young, R. M., & Howes, A. (1996). A dual-space model of iteratively deepening exploratory learning. *International Journal of Human-Computer Studies, 44*, 743–775.

Royer, C., Farahat, A. O., Pirolli, P., Budiu, R. (2005). GLSA Server @PARC. Proceedings of the *Twelfth Annual ACT-R Workshop*. Trieste, Italy.

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research, 1(4),* 201–220.

Salvucci, D. D., & Anderson, J. R. (2001). Automated eye-movement protocol analysis. *Human-Computer Interaction, 16*, 39–86.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26*, 1270-1281.

Sellen, A. J., Murphy, R., & Shaw, K. L. (2002). How knowledge workers use the web. In L. Terveen (Ed.), *Proceedings of the ACM CHI 2002 Human Factors in Computing Systems Conference* (pp. 227–234). New York, NY: ACM Press.

Snowberry, K., Parkinson, S. R., & Sisson, N. (1983). Computer display menus. *Ergonomics, 26*, 699–712.

Sohn, M.-H., Anderson, J. R., Reder, L. M., & Goode, A. (2004). Differential fan effect and attentional focus. *Psychonomic Bulletin and Review11(4)*, 729–734.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt, & P. A. Flach (Eds.) Proceedings of the *12th European Conference on Machine Learning* (Vol. 2167, pp. 491–502), London, UK: Springer-Verlag.

Young, R.M. (1998). Rational analysis of exploratory choice. In M.Oaksford & N.Chater (Eds.). *Rational Models of Cognition.* Oxford: Oxford University Press.

FIGURE CAPTIONS

Figure 1. **(a) Typical eye movement fixation trace from Experiment 1. The goal statement for this menu was "*Find a road map of Cardiff*", and the second item in the menu "*City Maps*" was the target. Rectangular boxes around menu items define areas of interest (see procedure in Experiment 1 for details). (b) Schematic representation of the collapsed gaze sequence. Note that item gazes represent consecutive eye movement fixations to the same item as a single data point.**

Figure 2. **The mean number of items visited and revisited up to the initial selection of an item for Experiment 1. Error bars represent standard error of the mean.**

Figure 3. **Distribution of the number of items visited after the initial visit to the selected item for Experiment 1.**

Figure 4. **The mean duration of item gazes for Experiment 1. Error bars represent standard error of the mean.**

Figure 5. **The proportion of gaze transitions that were between spatially non-contiguous items for Experiment 1. Error bars represent standard error of the mean.**

Figure 6. **The proportion of trials in which the participants selected the target item after visiting it for the first time for Experiment 2. Error bars represent standard error of the mean.**

Figure 7. **The frequency with which each item in the menu was visited at least once for Experiment 2.**

FIGURES

Figure 1. **(a) Typical eye movement fixation trace from Experiment 1. The goal statement for this menu was "*Find a road map of Cardiff*", and the second item in the menu "*City Maps*" was the target. Rectangular boxes around menu items define areas of interest (see procedure in Experiment 1 for details). (b) Schematic representation of the collapsed gaze sequence. Note that item gazes represent consecutive eye movement fixations to the same item as a single data point.**

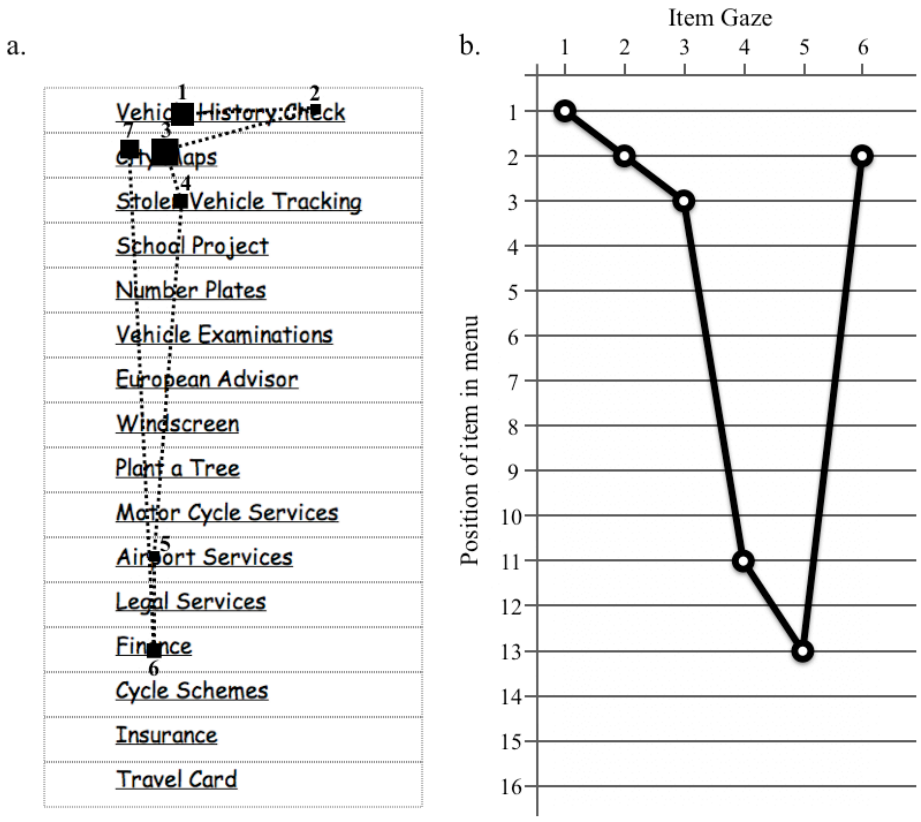

Figure 2. The mean number of items visited and revisited up to the initial selection of an item for Experiment 1. Error bars represent standard error of the mean.

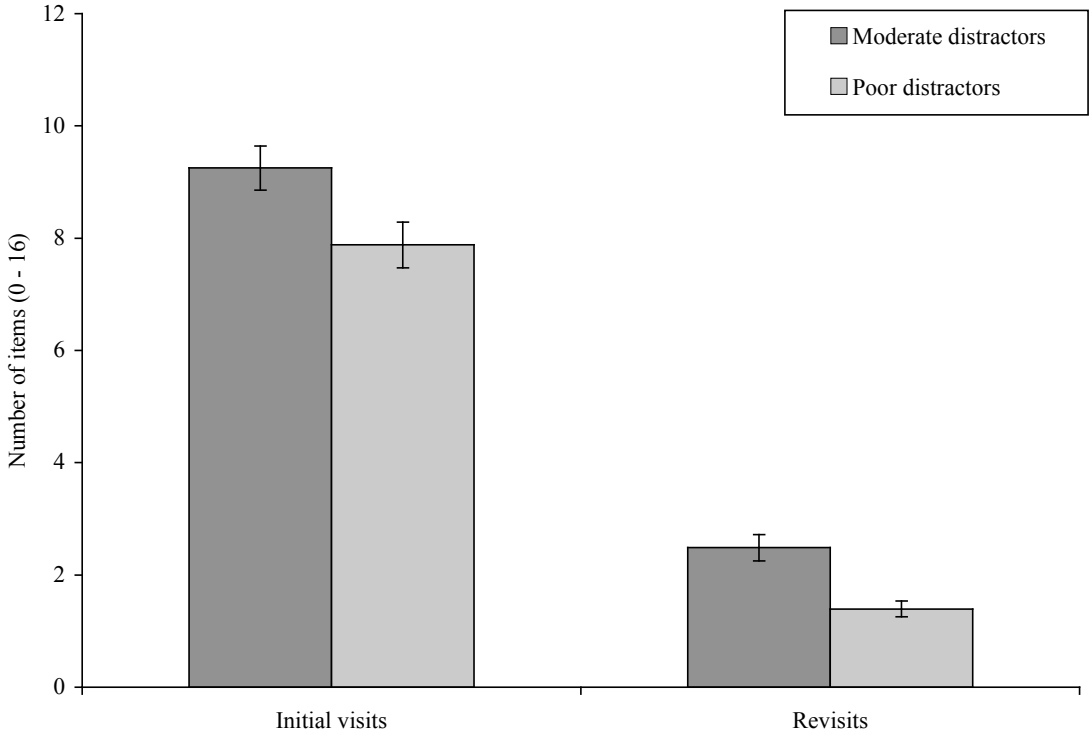

Figure 3. Distribution of the number of items visited after the initial visit to the selected item for Experiment 1

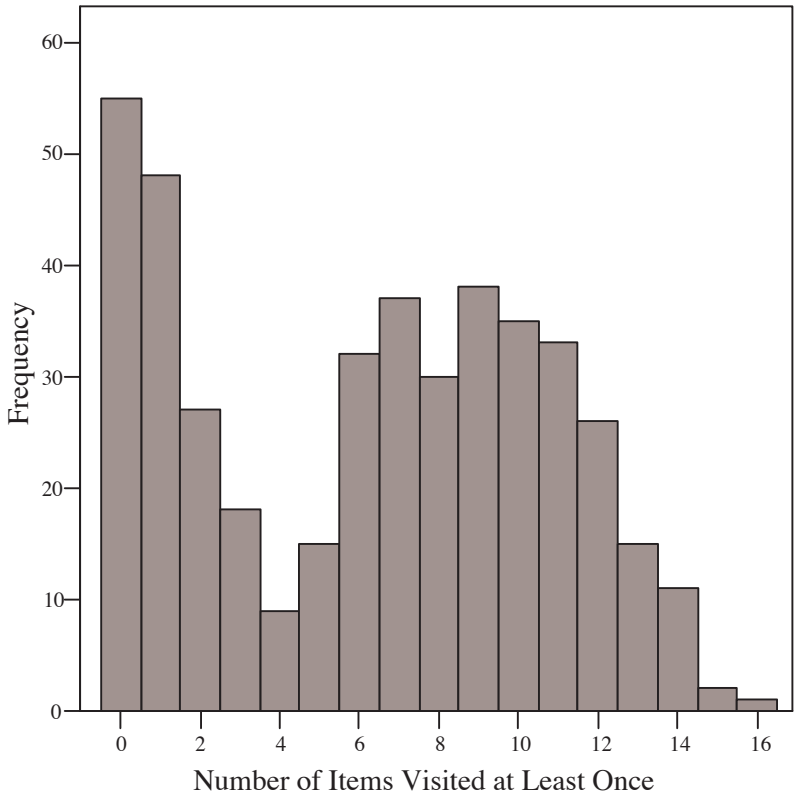

Figure 4. The mean duration of item gazes for Experiment 1. Error bars represent standard error of the mean.

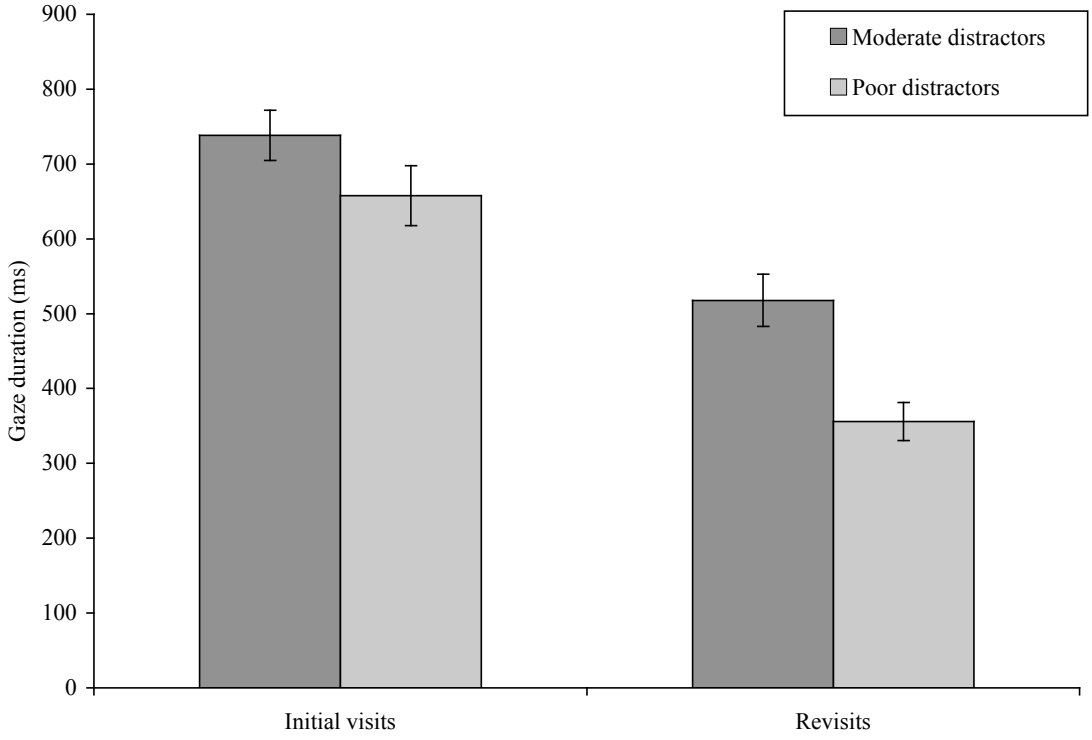

Figure 5. The proportion of gaze transitions that were between spatially non-contiguous items for Experiment 1. Error bars represent standard error of the mean.

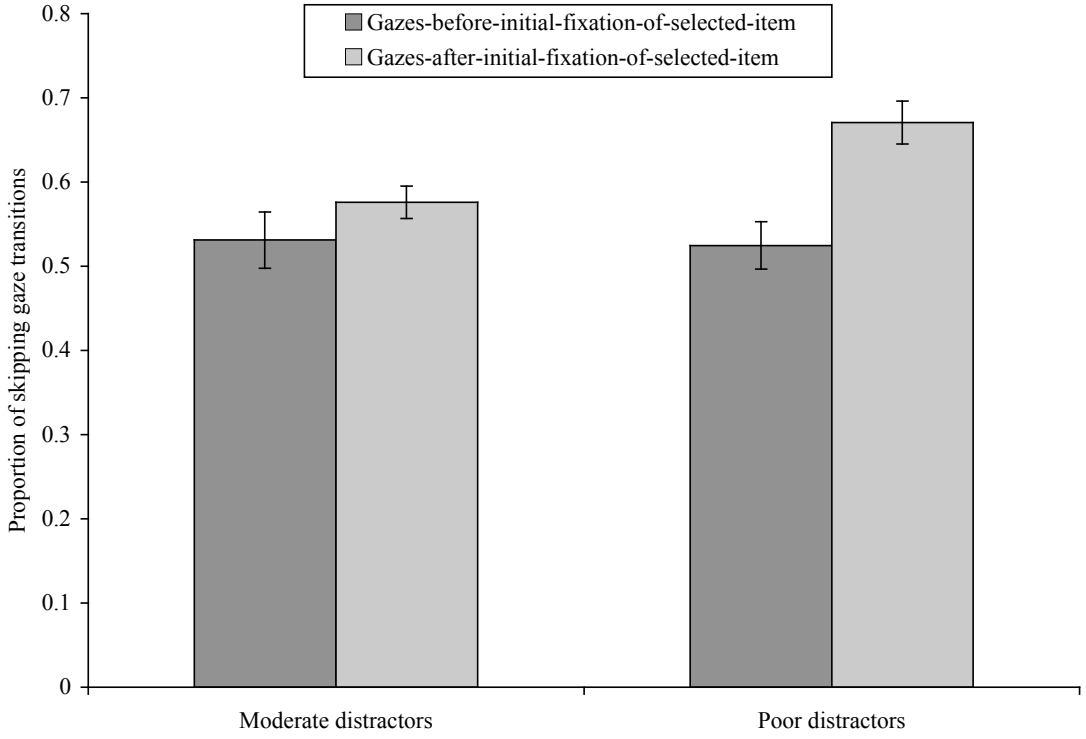

Figure 6. The proportion of trials in which the participants selected the target item after visiting it for the first time for Experiment 2. Error bars represent standard error of the mean.

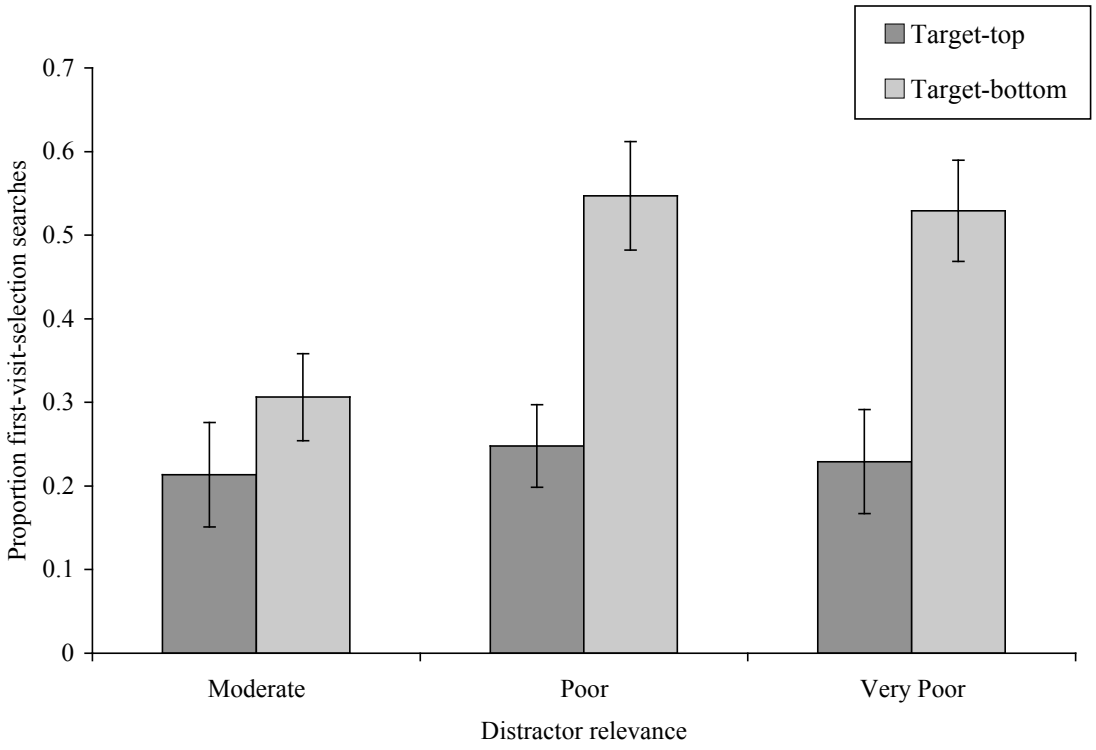

Figure 7. The frequency with which each item in the menu was visited at least once for Experiment 2.